RICE UNIVERSITY

Improved Biomolecular Crystallography at Low Resolution with the Deformable Complex Network Approach

by

Chong Zhang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE

Jianpeng Ma, Chair, Professor of Bioengineering, Rice University and Lodwick T. Bolin Professor of Biochemistry, Baylor College of Medicine

Huey W. Huang, Sam and Helen Worden Chair Professor of Physics and Astronomy, Rice University

Robert Raphael, Associate Professor of Bioengineering, Rice University

HOUSTON, TEXAS
SEPTEMBER 2012

ABSTRACT

Improved Biomolecular Crystallography at Low Resolution with the Deformable Complex Network Approach

by

Chong Zhang

It is often a challenge to atomically determine the structure of large macromolecular assemblies, even if successfully crystallized, due to their weak diffraction of X-rays. Refinement algorithms that work with low-resolution diffraction data are necessary for researchers to obtain a picture of the structure from limited experimental information. Relationship between the structure and function of proteins implies that a refinement approach delivering accurate structures could considerably facilitate further research on their function and other related applications such as drug design.

Here a refinement algorithm called the Deformable Complex Network is presented. Computation results revealed that, significant improvement was observed over the conventional refinement and DEN refinement, across a wide range of test systems from the Protein Data Bank, indicated by multiple criteria, including the free R value, the Ramachandran Statistics, the GDT (<1Å) score, TM-score as well as associated electron density map.

Acknowledgments

I thank my department, advisor and my fellow students for support.

I appreciate committee members Dr. Huang and Dr. Raphael for advice on revision of this thesis.

Computing resources at Rice University are acknowledged.

Contents

List of Figures

List of Tables

List of Equations

Chapter 1

Background

Structural determination is becoming more important and challenging as biomolecules grow more complicated. Investigation of relationship between structure and function in biological molecules urgently requires the development of advanced theoretic and computational technique capable of handling various systems with limited experiment data. A good refinement algorithm that refines a 3D-structure to a high accuracy could considerably facilitate the research on its functions, which ultimately leads to breakthroughs in other meaningful application areas including drug design.

1.1. X-ray and diffraction

X-ray diffraction is the most widely used method for biomolecular determination, others including nuclear magnetic resonance (NMR), electron microscopy, etc. Here we only focus on introduction to principles of X-ray diffraction and algorithm that deals with experiment data collected by X-ray equipments.

1.1.1. Property of X-ray

X-ray can be produced through an X-ray tube, which utilizes high voltage to accelerate electrons, make them hit a metal target and release X-rays due to energy level transition. The photon X-ray is also a form of electromagnetic wave, whose wavelength typically falls approximately between 0.01nm - 10nm. This feature makes X-ray the ideal tool for detecting atomic structures, in that the wavelength is comparable with crystal lattice and incident beam is easily diffracted. Conventionally, we denote the incident wave vector and wavelength with $\vec{k}$ and λ, respectively. The direction of $\vec{k}$ indicates the propagation of the X-ray, while the magnitude is defined to satisfy the following condition

$$\left|\vec{k}\right| = \frac{2\pi}{\lambda}$$

Equation 1-1 Relationship between X-ray wave vector and wavelength

1.1.2. Bragg's Law and diffraction

When crystals with periodic arrangement of atomic structures are bombarded by X-ray, reflected X-ray beams can be observed at certain incidence angles. The following graph[1] illustrates the situation in which two rays in parallel and in phase bombard two layers of a crystal. The dots within the crystal may represent atoms, icons and even molecules (including biological macromolecules). d denotes the distance between the two layers and θ the angle between the incident beam and crystal layer plane. We let $\vec{k}'$ be the scattered X-ray wave vector.

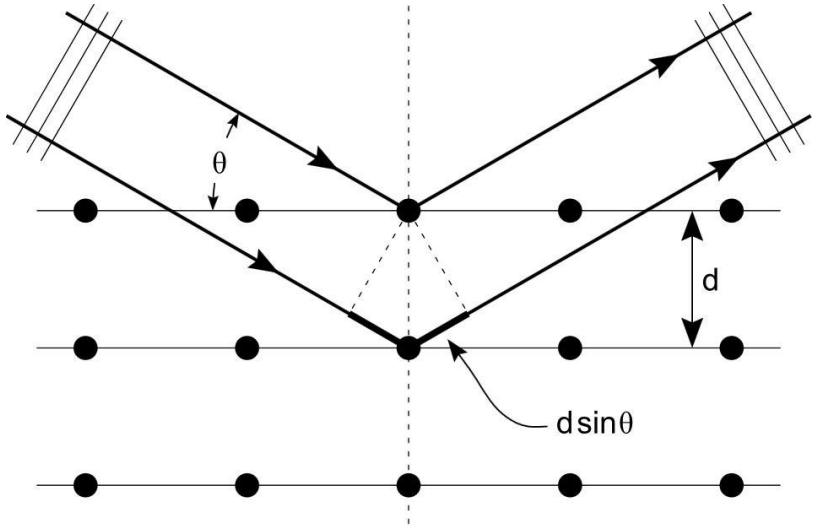


Figure 1-1 Bragg Law and Diffraction

Suppose that the scattering is elastic. Therefore, the X-ray photon energy – or the wavelength in terms of wave -- conserves before and after being diffracted by the lattice grids. In order to produce a constructive interference, the two outbound rays have to be in phase as well. It implies an identity stating that the optical path difference of these two rays is an integer multiple of their wavelength.

$$2d\sin\theta = n\lambda \qquad n = 1,2,3...$$

Equation 1-2 Bragg's Law

Bragg's Law describes the simplest case of diffraction. There exist other statements such as Lauer condition, which could be proven equivalent. The equation shows that by rotating the ray/crystal and changing the incident angle, diffractions patterns of higher orders, referred to as Bragg peaks, are to be observed. This idea forms

the principle of modern X-ray crystallography, where lattice grids become larger and more complex. By studying experiment patterns diffracted by target crystal during continuous orientation change, researchers are able to decipher the exact 3D-structure of a biomolecule, with the aid of powerful computation resources, advanced mathematical algorithm and sometimes other *a priori* information.

1.1.3. Diffraction of molecular crystal and structure factor

Suppose that we have a monoatomic crystal with a basis in hand. Two atoms within the same cell positioning at $\vec{R}_1$ and $\vec{R}_2$ will scatter the X-ray with a phase difference that equals

$$\Delta\phi_{1,2} = (\vec{R}_1 - \vec{R}_2)\cdot\vec{h}$$

Equation 1-3 Phase difference between two atoms in a crystal

Where $\vec{h}$ is a Bragg peak and defined as the wave vector difference between the scattered and incident X-rays

$$\vec{h} \equiv \vec{k}' - \vec{k}$$

Equation 1-4 definition of wave vector difference

As Equation 1-3 holds for all atoms within the crystal, the amplitude of the wave scattered by atom i and j will differ by a factor $e^{i\vec{h}\cdot(\vec{R}_i-\vec{R}_j)}$. Thus, scattered X-ray at each

position $\vec{R}_i$ should be proportional to $e^{\mathrm{i}\vec{h}\cdot\vec{R}_i}$. The final amplitude scattered by a primitive cell would be the total of individual ones

$$\mathbf{S}(\vec{h}) = \sum_{i=1}^{cell} e^{\mathrm{i}\vec{h}\cdot\vec{R}_i}$$

Equation 1-5 Structure factor for a monoatomic crystal with a basis

Where the summation is taken over all atoms belong to the same primitive cell. $S(\vec{h})$ is usually called the structure factor, which depends on the internal 3D arrangement of atoms for a lattice grid. The overall diffraction amplitude shall be accounted for by atoms within the entire crystal, and reduced using the structure factor as

$$\begin{aligned}\mathbf{F}(\vec{h}) &\propto \sum_{k=1}^{crystal} e^{\mathrm{i}\vec{h}\cdot\vec{r_k}} = \sum_{n_1,n_2,n_3,i} e^{\mathrm{i}\vec{h}\cdot(n_1\vec{a}+n_2\vec{b}+n_3\vec{c}+\vec{R_i})} \\ &= \sum_{n_1} e^{\mathrm{i}\vec{h}\cdot n_1\vec{a}} \sum_{n_2} e^{\mathrm{i}\vec{h}\cdot n_2\vec{b}} \sum_{n_3} e^{\mathrm{i}\vec{h}\cdot n_3\vec{c}} \sum_{i} e^{\mathrm{i}\vec{h}\cdot\vec{R_i}} \\ &= \mathbf{M}_F \cdot \mathbf{S}(\vec{h})\end{aligned}$$

Equation 1-6 Diffraction amplitude expression in terms of structure factor

Here, $\vec{r_k}$ is the atom position with respect to crystal origin, while $\vec{R_i}$ to lattice origin. $\vec{a},\vec{b},\vec{c}$ are three crystal lattice parameters. Assuming the crystal has N_1, N_2, N_3 lattice grids along $\vec{a},\vec{b},\vec{c}$ directions, respectively. Since the intensity is the square of the amplitude, we can write down

$$I(\vec{h}) = F(\vec{h})^2 \propto M_F^2 \cdot S(\vec{h})^2$$

Equation 1-7 Intensity, interference coefficient and structure factor

M_F^2 is sometimes referred to as the interference coefficient, which can be reduced since those three summations are actually geometric sequences

$$\begin{aligned} M_F^2 &= \left(\sum_{n_1=0}^{N_1-1} e^{\mathrm{i} n_1 \vec{h}\cdot\vec{a}} \sum_{n_2=0}^{N_2-1} e^{\mathrm{i} n_2 \vec{h}\cdot\vec{b}} \sum_{n_3=0}^{N_3-1} e^{\mathrm{i} n_3 \vec{h}\cdot\vec{c}} \right)^2 \\ &= \left(\frac{e^{\mathrm{i} N_1 \vec{h}\cdot\vec{a}} - 1}{e^{\mathrm{i}\vec{h}\cdot\vec{a}} - 1} \cdot \frac{e^{\mathrm{i} N_2 \vec{h}\cdot\vec{b}} - 1}{e^{\mathrm{i}\vec{h}\cdot\vec{b}} - 1} \cdot \frac{e^{\mathrm{i} N_3 \vec{h}\cdot\vec{c}} - 1}{e^{\mathrm{i}\vec{h}\cdot\vec{c}} - 1} \right)^2 \\ &= \frac{\sin^2\left(\frac{N_1 \vec{h}\cdot\vec{a}}{2}\right)}{\sin^2\left(\frac{\vec{h}\cdot\vec{a}}{2}\right)} \cdot \frac{\sin^2\left(\frac{N_2 \vec{h}\cdot\vec{b}}{2}\right)}{\sin^2\left(\frac{\vec{h}\cdot\vec{b}}{2}\right)} \cdot \frac{\sin^2\left(\frac{N_3 \vec{h}\cdot\vec{c}}{2}\right)}{\sin^2\left(\frac{\vec{h}\cdot\vec{c}}{2}\right)} \end{aligned}$$

Equation 1-8 Reducing the interference coefficient

The characteristics of the function is that, the maximum value $N_1^2 N_2^2 N_3^2$ is reached whenever the following condition is met

$$\begin{cases} \vec{a}\cdot\vec{h} = m_1 \cdot 2\pi \\ \vec{b}\cdot\vec{h} = m_2 \cdot 2\pi \\ \vec{c}\cdot\vec{h} = m_3 \cdot 2\pi \end{cases} \qquad m_1, m_2, m_3 = 1, 2, 3...$$

Equation 1-9 Condition when I_F reaches maximum

It is easy to show that Equation 1-9 is equivalent to Lauer condition and Bragg Law Equation 1-2 along three different crystal axes. Note, the interference coefficient also maximizes when m_1, m_2, m_3 equal zero, in which case the incident X-ray is parallel to one of the crystal planes from the view of Figure 1-1.

M_F^2 also has multiple other local maxima for other m_1, m_2, m_3 values with lower peaks. However, at other m_1, m_2, m_3 configurations, M_F^2 quickly vanishes. The larger N_1, N_2, N_3 are, the more quickly M_F^2 vanishes between those maxima, the closer they are apart from each other, and the larger their peak height difference will be. The behavior is shown as follows with increasing N .

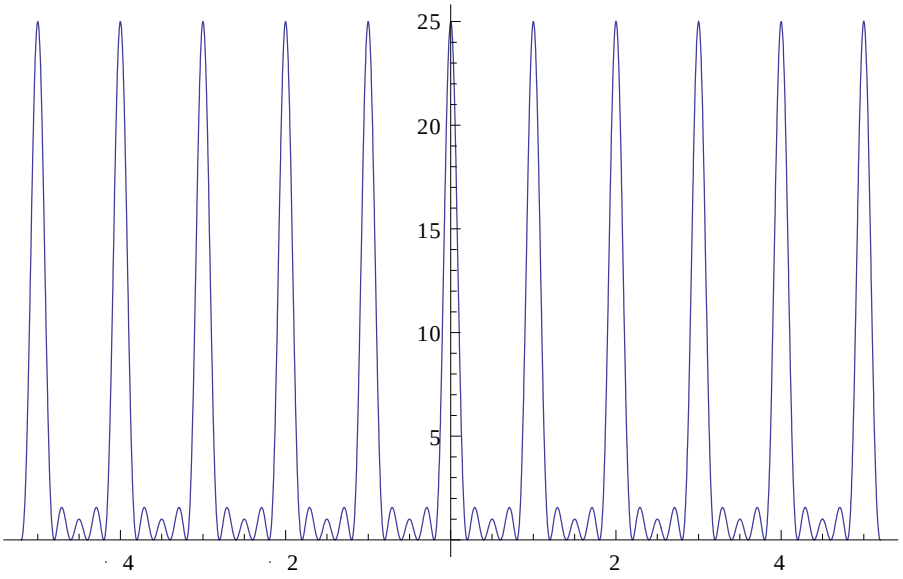


Figure 1-2 Plot $f(x) = \dfrac{\sin^2(N\pi \cdot x)}{\sin^2(\pi \cdot x)}$ $\qquad N = 5$

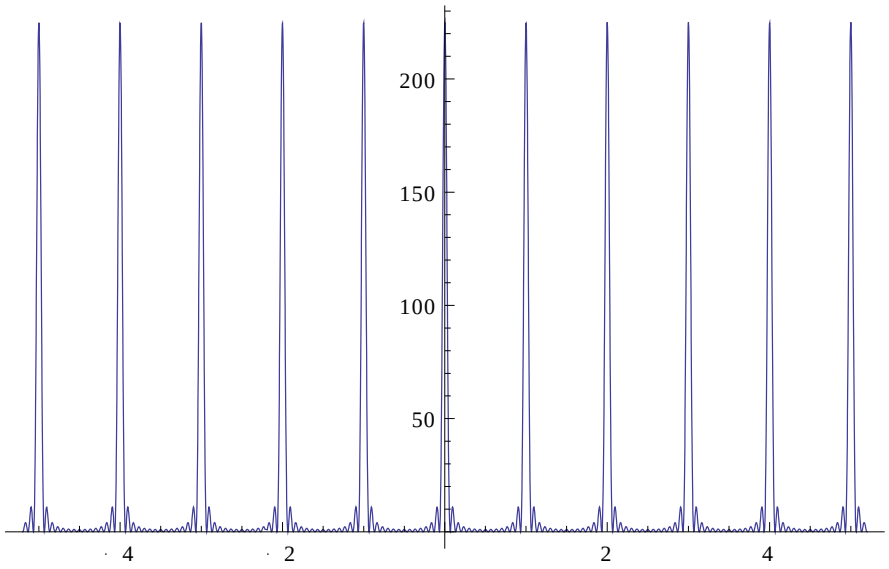


Figure 1-3 Plot $f(x)=\dfrac{\sin^2(N\pi\cdot x)}{\sin^2(\pi\cdot x)}$ $N=15$

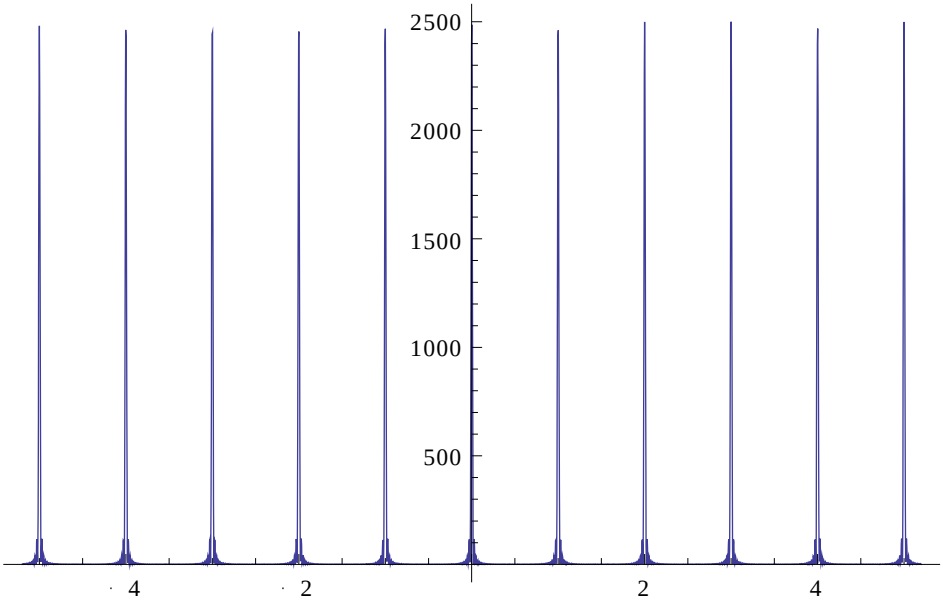


Figure 1-4 Plot $f(x)=\dfrac{\sin^2(N\pi\cdot x)}{\sin^2(\pi\cdot x)}$ $N=50$

This makes the actual $\vec{h}$ dependence of $M_F^2(\vec{h})$ less sensitive and distinctive than $S(\vec{h})$. The profile of structure factors is derived simply after some scaling operations from that of the experiment amplitudes $F(\vec{h})$.

As for molecular crystals (or polyatomic crystals) whose primitive cell contains multiple species, the diffraction ability of each atom has to be distinguished and appropriately weighted. Therefore, we take the structure factor

$$\mathbf{S}(\vec{h}) = \sum_{i=1}^{cell} f_i(\vec{h}) e^{\mathrm{i}\vec{h}\cdot\vec{R}_i}$$

Equation 1-10 structure factor for molecular crystal

Where $f_i(\vec{h})$ is the atomic form factor. It depends on the wave vector difference and, more importantly, the electron density distribution $\rho(\vec{r})$ of an atom, since X-rays bombarded on to the crystal are actually interacting with and diffracted by electrons of each atom.

$$f_i(\vec{h}) = \int_{atom} \rho_i(\vec{r'}) \cdot e^{\mathrm{i}\vec{h}\cdot\vec{r'}} \,\mathrm{d}\vec{r'}$$

Equation 1-11 atomic form factor and electron density

$\vec{r'}$ is the position of electron density within a specified atom. In fact, taking elementary treatment, we can combine the summation over all atoms in a cell and the integral over electrons in an atom into an expanded integral, in which we count the

diffraction contribution directly from electrons but integrate across the entire unit cell space for the structure factor.

$$\mathbf{S}(\vec{h}) = \int_{cell} \rho(\vec{r}) \cdot e^{\mathrm{i}\vec{h}\cdot\vec{r}} \, \mathrm{d}\vec{r}$$

Equation 1-12 structure factor and electron density

It is straightforward to obtain the 3D electron density information, i.e., the molecular structure, via a Fourier Transform of Equation 1-10

$$\rho(\vec{r}) = \frac{1}{(2\pi)^3} \int_{diffractions} \mathbf{S}(\vec{h}) \cdot e^{-\mathrm{i}\vec{h}\cdot\vec{r}} \, \mathrm{d}\vec{h}$$

Equation 1-13 electron density and structure factor

Experiment data collected on a screen are X-ray intensities. They are proportional to the square of the X-ray amplitudes, which are subsequently proportional to the square of the structure factors. Since structure factors are complex, by squaring them we are in essence multiplying them with their conjugates. This leads to a data loss – all phase information is wiped out during this process with only the magnitude left, making it impossible to solve structure problem by immediate adoption of Equation 1-10.

1.2. Basics of biomolecular structure determination and refinement

1.2.1. Crystallographic structure determination of biomolecules

The nature of measurement inhibits the detection of diffraction phases. What we can do, however, is to try to estimate an ‘initial phase’ (herein after ‘phasing’) that is believed to have commonality with and close to that of the target structure, through various experimental or theoretic techniques. We then build a rough model based on the resultant electron density map, derived from the initial phases and experiment intensities (amplitudes). Now that the atoms’ positions are known, by going back to use the original experiment amplitudes it is easy to calculate a more reliable density map. Those current atomic coordinates are again adjusted towards the newly updated density map. By continuously changing the coordinates and other parameters of atoms (referred to as ‘refinement’) in order to fit the self-produced map or experiment data and optimize a target function, the structure of the molecule is gradually improved and more self-consistent. This procedure proceeds until the agreement between the structure and data converges to a satisfactory level, generally indicated by the R value.

When coming to determine a specific category of objects, that is, the biomolecules, it is obvious that some experiment procedures have to be involved before beginning data collection. The following chart illustrates the complete workflow, from purifying studies sample to publishing the determined structure and depositing into certain commonly accessible online database (such as the Protein Data Bank).

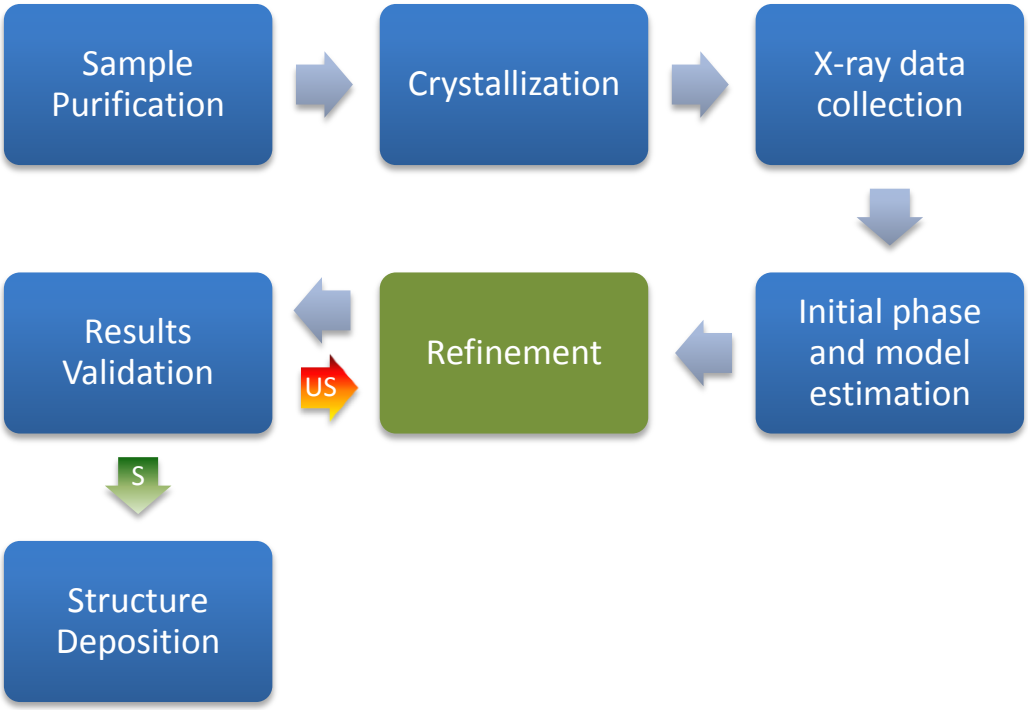


Figure 1-5 Workflow of crystallographic biomolecular structure determination

Should results after refinement are determined unsatisfactory by several validation criteria, the current structure is sent back to be refined again. The process repeats until the final structure is satisfactory for publishing and deposition.

1.2.2. Structure refinement

Refinement is an iterated process where the molecular parameters are constantly changed in order to optimize a target function that takes experiment data, stereochemistry and other factors into consideration. When refinement proceeds, structure of the molecule

is improved and able to better describe the diffraction data, and finally yield an accurate model.

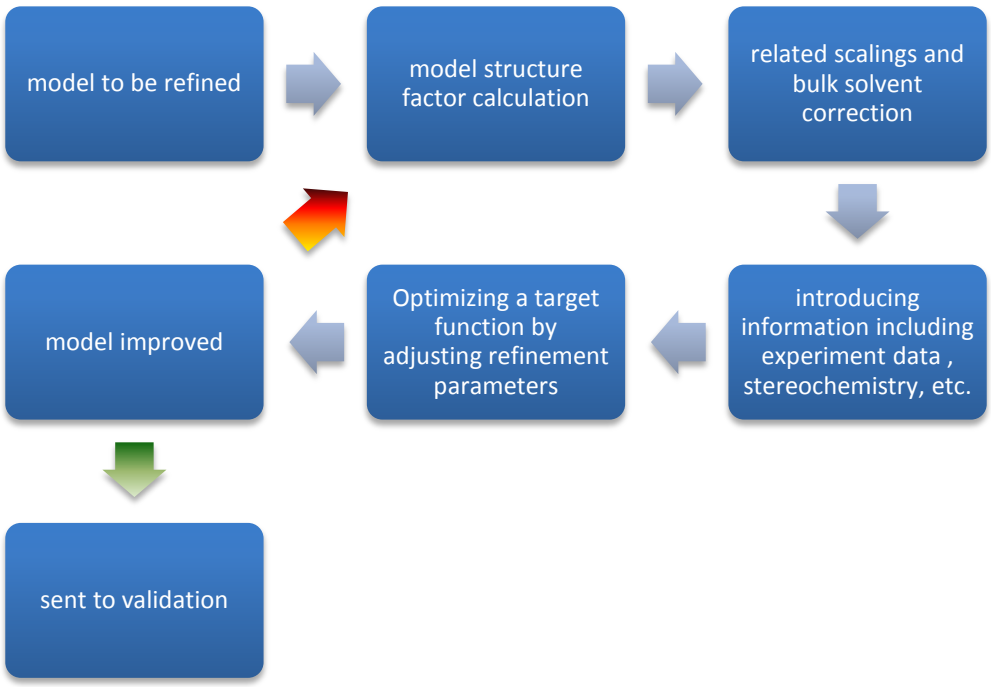


Figure 1-6 Work flow of refinement

Usually a whole refinement is subject to multiple macro-cycles, where the improved model after a cycle serves as 'to-be-refined' and is immediately sent to a new round. It is necessary due to the complexity of the energy landscape profile of a biomolecule with a large number of parameters.

In addition to the application in solving unknown structures, refinement is also useful for improving structures of already deposited molecules. There are possibilities to do so, as well motivations. The more precisely a molecular structure is determined, the

more smoothly its function shall be researched, and the more promising it is exploited in various application fields.

Chapter 2

Refinement theories and techniques

2.1. Refinement parameters

The aim of refinement is to modify model parameters to improve the overall structure. What are those parameters? Clearly, the coordinates of each atom within a molecule is one of the most important types of parameters. Other parameters that can be refined include the B-factor and atom occupancy. These are due to the local thermal fluctuation of atoms, disorder in crystals and other sorts of molecular motions.

2.1.1. Atomic coordinates

The position of atoms is described via a set of three dimensional coordinates, typically in the Cartesian coordination system. The main task of refinement is to move the coordinates of every atom, go through a wide range of conformation change and energy landscape and finally improve the structure by best explaining the experiment data. The number of parameters per atom necessary to express the coordinates is three ---

(x, y, z). In accordance with the PDB format, the three coordinates are listed at columns 31 to 54 in a PDB file[2], which is colored red below.

Example:

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM    145  N   VAL A  25      32.433  16.336  57.540  1.00 11.92      A1   N
ATOM    146  CA  VAL A  25      31.132  16.439  58.160  1.00 11.85      A1   C
ATOM    147  C   VAL A  25      30.447  15.105  58.363  1.00 12.34      A1   C
ATOM    148  O   VAL A  25      29.520  15.059  59.174  1.00 15.65      A1   O
ATOM    149  CB AVAL A  25      30.385  17.437  57.230  0.28 13.88      A1   C
ATOM    150  CB BVAL A  25      30.166  17.399  57.373  0.72 15.41      A1   C
ATOM    151  CG1AVAL A  25      28.870  17.401  57.336  0.28 12.64      A1   C
ATOM    152  CG1BVAL A  25      30.805  18.788  57.449  0.72 15.11      A1   C
ATOM    153  CG2AVAL A  25      30.835  18.826  57.661  0.28 13.58      A1   C
ATOM    154  CG2BVAL A  25      29.909  16.996  55.922  0.72 13.25      A1   C
```

Table 2-1 Example of a PDB file[2]

2.1.2. B-factor

B-factor accounts for the local mobility of an atom around its equilibrium position. Historically, B-factor was also called the thermal factor or Debye-Waller factor.

Assume the instantaneous position of j th atom in a unit cell $\vec{r}_j$

$$\vec{r}_j = \left\langle \vec{r}_j \right\rangle + \Delta\vec{r}_j$$

Equation 2-1 Instantaneous atom position

Where $\left\langle \vec{r}_j \right\rangle$ is the equilibrium position of atom j and $\Delta\vec{r}_j$ the deviation from that position. Experiment measurement is always a time-averaged ensemble of numerous instant conformations of the molecule. Therefore, the structure factor should be an average of the previous definition.

$$S(\vec{h}) = \left\langle \sum_{j=1}^{cell} f_j(\vec{h}) e^{\mathrm{i}\vec{h}\cdot\vec{r}_j} \right\rangle = \sum_{j=1}^{cell} f_j(\vec{h}) \left\langle e^{\mathrm{i}\vec{h}\cdot\vec{r}_j} \right\rangle$$

Equation 2-2 Structure factor time averaged

Insert Equation 2-1into the last factor of Equation 2-2,

$$\begin{aligned}
&\left\langle e^{\mathrm{i}\vec{h}\cdot\vec{r}_j} \right\rangle = \left\langle e^{\mathrm{i}\vec{h}\cdot\left(\langle\vec{r}_j\rangle+\Delta\vec{r}_j\right)} \right\rangle = e^{\mathrm{i}\vec{h}\cdot\langle\vec{r}_j\rangle} \cdot \left\langle e^{\mathrm{i}\vec{h}\cdot\Delta\vec{r}_j} \right\rangle \\
&= e^{\mathrm{i}\vec{h}\cdot\langle\vec{r}_j\rangle} \cdot \left\langle 1 + \mathrm{i}\vec{h}\cdot\Delta\vec{r}_j + \frac{1}{2!}\left(\mathrm{i}\vec{h}\cdot\Delta\vec{r}_j\right)^2 + \cdots \right\rangle \\
&= e^{\mathrm{i}\vec{h}\cdot\langle\vec{r}_j\rangle} \cdot \left(1 + \mathrm{i}\vec{h}\cdot\left\langle\Delta\vec{r}_j\right\rangle - \frac{1}{2}\left\langle\left(\vec{h}\cdot\Delta\vec{r}_j\right)^2\right\rangle + \cdots \right) \\
&= e^{\mathrm{i}\vec{h}\cdot\langle\vec{r}_j\rangle} \cdot \left(1 - \frac{1}{2}\left\langle\left(\vec{h}\cdot\Delta\vec{r}_j\right)^2\right\rangle + \cdots \right) \\
&\approx e^{\mathrm{i}\vec{h}\cdot\langle\vec{r}_j\rangle} \cdot e^{-\frac{1}{2}\left\langle\left(\vec{h}\cdot\Delta\vec{r}_j\right)^2\right\rangle}
\end{aligned}$$

Equation 2-3 Reducing $\left\langle e^{\mathrm{i}\vec{h}\cdot\vec{r}_j} \right\rangle$

Here we used $\left\langle\Delta\vec{r}_j\right\rangle = 0$ due to $\left\langle\vec{r}_j\right\rangle = \left\langle\left\langle\vec{r}_j\right\rangle + \Delta\vec{r}_j\right\rangle = \left\langle\vec{r}_j\right\rangle + \left\langle\Delta\vec{r}_j\right\rangle$. The last step is a Gaussian Approximation.

We can write the exponential factor $-\frac{1}{2}\left\langle\left(\vec{h}\cdot\Delta\vec{r_j}\right)^2\right\rangle$ in the form of matrix U_j and row vector $\vec{h}$.

$$
\begin{aligned}
&-\frac{1}{2}\left\langle\left(\vec{h}\cdot\Delta\vec{r_j}\right)^2\right\rangle \\
&=-\frac{1}{2}\left\langle\left(\vec{h}\cdot\Delta\vec{r_j}\right)\cdot\left(\Delta\vec{r_j}\cdot\vec{h}\right)\right\rangle \\
&=-\frac{1}{2}\left\langle\vec{h}\left(\Delta\vec{r_j}\times\Delta\vec{r_j}\right)\vec{h}^T\right\rangle \\
&=-\frac{1}{2}\vec{h}\left\langle\Delta\vec{r_j}\times\Delta\vec{r_j}\right\rangle\vec{h}^T \\
&\equiv-\frac{1}{2}\vec{h}U_j\vec{h}^T
\end{aligned}
$$

Equation 2-4 U matrix

With

$$
U_j\equiv\left\langle\Delta\vec{r_j}\times\Delta\vec{r_j}\right\rangle=\begin{pmatrix}\left\langle\Delta x_j\Delta x_j\right\rangle & \left\langle\Delta x_j\Delta y_j\right\rangle & \left\langle\Delta x_j\Delta z_j\right\rangle\\ \left\langle\Delta y_j\Delta x_j\right\rangle & \left\langle\Delta y_j\Delta y_j\right\rangle & \left\langle\Delta y_j\Delta z_j\right\rangle\\ \left\langle\Delta z_j\Delta x_j\right\rangle & \left\langle\Delta z_j\Delta y_j\right\rangle & \left\langle\Delta z_j\Delta z_j\right\rangle\end{pmatrix}
$$

Equation 2-5 U matrix definition

$\Delta x_j,\Delta y_j,\Delta z_j$ are the three component of $\Delta\vec{r_j}$. If the vibration is approximated to be harmonic and isotropic, then

$$U_j = \begin{pmatrix} \frac{\Delta r_j^2}{3} & & \\ & \frac{\Delta r_j^2}{3} & \\ & & \frac{\Delta r_j^2}{3} \end{pmatrix}$$

$$-\frac{1}{2}\left\langle \left(\vec{h}\cdot\Delta\vec{r_j}\right)^2 \right\rangle = -\frac{1}{6}\vec{h}^2\cdot\Delta r_j^2 \equiv -\frac{1}{16\pi^2}\vec{h}^2\cdot B_j$$

Equation 2-6 Introducing B factor

Here shows the isotropic B-factor definition

$$B_j \equiv \frac{8\pi^2}{3}\Delta r_j^2$$

Equation 2-7 Definition of Isotropic B factor

a quantity that directly proportional to the fluctuation of the atom. Δr_j^2 is the mean square of *total* displacement with respect to the equilibrium, not that of a specific direction. There exist other expressions where the diagonal term of U_j is denoted by u_j^2, which lead to a slightly different form to Equation 2-7

$$B_j \equiv 8\pi^2 u_j^2$$

Equation 2-8 Alternative form of definition of Isotropic B factor

The coefficient introduced in Equation 2-6 is for the purpose of simplifying the thermal term $\left\langle e^{i\vec{h}\cdot\Delta\vec{r}_j} \right\rangle$

$$\left\langle e^{i\vec{h}\cdot\Delta\vec{r}_j} \right\rangle = e^{-\frac{1}{16\pi^2}\vec{h}^2\cdot B_j} = e^{-B_j\left(\frac{\sin\theta}{\lambda}\right)^2}$$

Equation 2-9 Thermal term expressed by B factor, incident angle and wavelength

with the aid from the Lauer Condition

$$\vec{h}^2 = 2\cdot\sin\theta\cdot\frac{2\pi}{\lambda}$$

The Isotropic B-factor value is recorded at columns 61- 66 in a PDB file, colored in blue below.

Example:

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM    145  N   VAL A  25      32.433  16.336  57.540  1.00 11.92      A1   N
ATOM    146  CA  VAL A  25      31.132  16.439  58.160  1.00 11.85      A1   C
ATOM    147  C   VAL A  25      30.447  15.105  58.363  1.00 12.34      A1   C
ATOM    148  O   VAL A  25      29.520  15.059  59.174  1.00 15.65      A1   O
ATOM    149  CB AVAL A  25      30.385  17.437  57.230  0.28 13.88      A1   C
ATOM    150  CB BVAL A  25      30.166  17.399  57.373  0.72 15.41      A1   C
ATOM    151  CG1AVAL A  25      28.870  17.401  57.336  0.28 12.64      A1   C
ATOM    152  CG1BVAL A  25      30.805  18.788  57.449  0.72 15.11      A1   C
ATOM    153  CG2AVAL A  25      30.835  18.826  57.661  0.28 13.58      A1   C
ATOM    154  CG2BVAL A  25      29.909  16.996  55.922  0.72 13.25      A1   C
```

Table 2-1 Example of a PDB file2 [2]

In refinement, these B-factors for each atom are treated like independent parameters, though restraints on them are imposed from time to time.

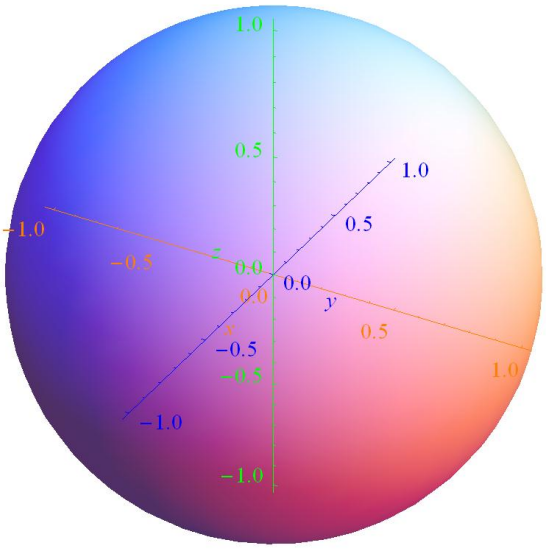


Figure 2-1 Isotropic oscillation shell—Sphere

When anisotropicity is considered, a total of 6 independent parameters, which correspond to the three diagonal terms and three non-diagonal terms of matrix U_j are used for each atom. In practice, a separate line starting with 'ANISOU' under each 'ATOM' entry is deployed. Columns 29-70 in each ANISOU entry are reserved for six anisotropic B-factor terms in the order of $B_{11}, B_{22}, B_{33}, B_{13}, B_{13}, B_{23}$, each occupies seven columns(colored green below).

```
Example

         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
```

```
ATOM    107  N   GLY A  13      12.681  37.302 -25.211 1.000 15.56           N
ANISOU  107  N   GLY A  13     2406   1892   1614    198    519   -328       N
ATOM    108  CA  GLY A  13      11.982  37.996 -26.241 1.000 16.92           C
ANISOU  108  CA  GLY A  13     2748   2004   1679    -21    155   -419       C
ATOM    109  C   GLY A  13      11.678  39.447 -26.008 1.000 15.73           C
ANISOU  109  C   GLY A  13     2555   1955   1468     87    357   -109       C
ATOM    110  O   GLY A  13      11.444  40.201 -26.971 1.000 20.93           O
ANISOU  110  O   GLY A  13     3837   2505   1611    164   -121    189       O
ATOM    111  N   ASN A  14      11.608  39.863 -24.755 1.000 13.68           N
ANISOU  111  N   ASN A  14     2059   1674   1462     27    244    -96       N
```

Table 2-2 Example of a PDB file with ANISOU entries[2]

Use of anisotropic B factor will lead to a six-fold parameter increase and is subject to the completeness and quality of experimental diffraction data.

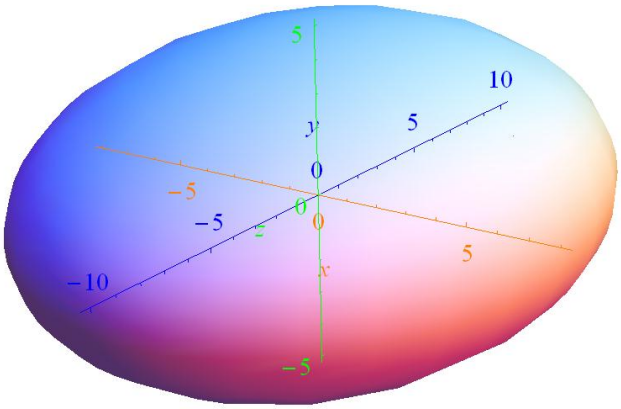


Figure 2-2 Anisotropic oscillation shell – Ellipsoid whose principle axes not necessarily x,y,z coordinate axes

2.1.3. Occupancy

Known as the static crystal disorder, all molecular copies from different region of the crystal are not exactly in identical conformation, due to the local flexibility of the protein side chains, or even main chains. Those alternative conformations can be stable and distinguishable by the electron density map (usually with high resolution data), exhibiting a significant population compare with the main conformation. In order to account for this, a parameter called atom occupancy is introduced to describe the possibility an atom is in a certain conformation state (position), or, from the statistical point of view, the ratio between the number of molecule copies where the atom is in the particular conformation and the total number of molecule copies within a crystal.

In case multiple conformations are detected and recorded, each occupancy is less than 1 and subject to refinement. One constraint is that the occupancy must add up to 1 for multiple conformations, except for those non-protein atoms, such as metal atom and other ligands, whose occupancy is allowed to be partial owing to imperfection in co-crystallization under certain experiment conditions. Occupancy of each atom is recorded at columns 55-60 in a PDB file, colored yellow below.

Example:

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM    145  N   VAL A  25      32.433  16.336  57.540  1.00 11.92      A1   N
ATOM    146  CA  VAL A  25      31.132  16.439  58.160  1.00 11.85      A1   C
ATOM    147  C   VAL A  25      30.447  15.105  58.363  1.00 12.34      A1   C
ATOM    148  O   VAL A  25      29.520  15.059  59.174  1.00 15.65      A1   O
ATOM    149  CB AVAL A  25      30.385  17.437  57.230  0.28 13.88      A1   C
ATOM    150  CB BVAL A  25      30.166  17.399  57.373  0.72 15.41      A1   C
ATOM    151  CG1AVAL A  25      28.870  17.401  57.336  0.28 12.64      A1   C
ATOM    152  CG1BVAL A  25      30.805  18.788  57.449  0.72 15.11      A1   C
ATOM    153  CG2AVAL A  25      30.835  18.826  57.661  0.28 13.58      A1   C
ATOM    154  CG2BVAL A  25      29.909  16.996  55.922  0.72 13.25      A1   C
```

Table 2-3 Example of a PDB file[2]

2.1.4. Bulk solvent correction and k_{sol}, B_{sol} optimization

There are additional two parameters, K_{sol}, B_{sol}, that are related to the bulk solvent. Bulk solvent is the region of the unit cell other than the protein molecule. The mask that separates molecule and the bulk solvent is called the solvent mask. Since solvent makes contribution to X-ray diffraction as well, it is important to account for it for a more accurate calculated structure factor to fit the experiment data.

One model to describe the bulk solvent is based on Babinet's Principle, which sits on the statement that a 180 (half wavelength) shift exists between the Fourier Transform of the solvent mask and the protein mask. The implication that the electron density of the solvent is proportional to that of the protein with opposite phases does not hold at resolution higher than $15A^3$, and therefore not recommended.

The other model is the flat density model, in which no assumption of structure factor relationship between the molecule and bulk solvent is made. This model needs to determine a relatively accurate mask by placing the molecule on a grid and distinguishing the grid points falling in and out of the molecular region, followed by further refining the boundary with two parameters SOLRAD and SHRINK, which are normally set as constant. k_{sol} represents the flat density of the solvent and another parameter B_{sol} is introduced to smooth sharp edge effects arising from the Fourier transform of grids between solvent and solvent-excluded regions[4]. The total structure is

$$\mathbf{S}_{tot}(\vec{h}) = \mathbf{S}(\vec{h})_{mol} + k_{sol} e^{-\frac{\vec{h}^2}{16\pi^2} \cdot B_{sol}} \mathbf{S}(\vec{h})_{sol}$$

or

$$\mathbf{S}_{tot}(\vec{h}) = \mathbf{S}(\vec{h})_{mol} + k_{sol} e^{-\frac{\sin^2\theta}{\lambda^2} \cdot B_{sol}} \mathbf{S}(\vec{h})_{sol}$$

Equation 2-10 Total calculated structure factor with flat density bulk solvent

These two parameters are usually optimized immediately followed by a new refinement macro-cycle, before other parameters begin refined. It is done by 'minimizing R value in lowest resolution shell without significantly increasing the high resolution R values[4], with parameters fixed one after the other in an iterative style in 2D space.

2.2. Refinement target

Refinement is a process that by optimizing a refinement target, parameters of a model are continuously changed and can better explain the experiment data, at the same time without generating irrational results judged by stereochemistry and *a priori* knowledge.

In a physics style, we treat the refinement target as a total potential energy term. The total energy should therefore involve experiment-related energy, stereochemistry energy and *a priori* knowledge based restraints potential. Our goal is to minimize an overall function linearly combined by all of them, rather than a particular one, across the function's complicated energy landscape.

$$E_{\text{target}} = f(E_{\text{experiment}}, E_{\text{stereo}}, E_{\textit{a priori knowledge}})$$

Equation 2-11 Refinement target, i.e., Total potential energy

2.2.1. Experiment-related energy

The ultimate gold of refinement is to best interpret the experiment data in hand by predicting an accurate model. The extent of agreement reached between model and data can be quantified as an energy term, and depends on the way experiment data are expressed.

2.2.1.1. Fitting electron density – real space refinement

The most original idea of refinement is to calculate experiment electron density map using experiment amplitude and model phase determined in the previous step, followed by fitting a current model density map to the experiment density, generating a new model and a new experiment density map, and repeat[5].

This way, the experiment-related energy term is defined as

$$E_{\text{experiment}} = \sum_{\text{grid points}} (\rho_{best} - k\rho_{calc})^2$$

Equation 2-12 experiment potential with electron density map

Where ρ_{best} is the best available map (an experiment map or $2mF_o - DF_c$ map), ρ_{calc} is model calculated map. Though this method was widely used before 1941[6], it is

seldom used in modern refinement algorithms, due to the dependence of the calculated density map on the quality of the diffraction data. Density map corresponding to identical structure but derived from data with different resolutions can differ quite much. This undesired feature makes real space refinement unreliable and gradually abandoned, especially after the emergence of those refinement targets in reciprocal space.

2.2.1.2. Least Squares target function– reciprocal space refinement

Hughes proposed a refinement target function in reciprocal space called Least Squares energy.

$$E_{\text{experiment}} = \sum_{\vec{h}} w(\vec{h})\left(F_{obs}(\vec{h}) - kF_{cal}(\vec{h})\right)^2$$

Equation 2-13 Least Squares target function

k is a scaling coefficient, F_{obs} and F_{cal} are observed and calculated diffraction amplitude, respectively. w is a weight to account for the importance of each diffraction's contribution to the total target function. The summation is taken over all diffractions. The least squares energy basically describes the agreement between the model calculated structure factors and their counterpart from diffraction data. It was widely used in small molecule refinement. The limitation of least squares target is that it is unable to smartly adjust the weight of contribution according to different quality of measurement of diffraction entries. Moreover, in cases where there are missing atoms or chemical groups in a model, least squares target fails to correctly interpret this situation and will subsequently guides the refinement towards unfavorable directions.

2.2.1.3. Maximum Likelihood target function

There is another way of expressing the agreement of the data and model, with ability to take incompleteness and error of the model into consideration[7], a desired feature for refining macromolecules.

The target function is defined as the likelihood of observing a data set (in this case the experiment data set in hand) given a known model. Our goal is to maximize this likelihood by modifying the model parameters, or minimize its opposite number (treated as a form of potential energy for a consistent style with other target function terms).

Assuming the conditional probability of an observation amplitude given the model is $P(F_{obs}(\vec{h}) \mid \mathbf{F}_{cal}(\vec{h}))$ and different diffraction entries are independent of each other. In order to observe a particular pattern of diffraction, all diffractions this pattern is composed of should be observed simultaneously. Therefore, the associated likelihood L is the multiplicity of the single diffraction's conditional probability.

$$L = \prod_{\vec{h}} P(F_{obs}(\vec{h}) \mid \mathbf{F}_{cal}(\vec{h}))$$

Equation 2-14 Likelihood of observing a diffraction pattern given a model

It is for computational convenience purpose that usually the logarithm of Equation 2-14 is taken to transform $\prod$ into $\sum$, and '-1' subsequently multiplied to achieve the purpose of maximizing Equation 2-14 by minimizing the following energy function:

$$E_{\text{experiment}} = -\ln L = -\sum_{\vec{h}} \ln P(F_{obs}(\vec{h}) \mid \mathbf{F}_{cal}(\vec{h}))$$

Equation 2-15 Maximum Likelihood target function definition

Accounting for model error σ_Δ, measurement error $P_{meas}(F_{obs} \mid F)$, *a priori* phase distribution P_{phase}, and re-write $\mathbf{F}(\vec{h})$ with $F(\vec{h}) \cdot e^{\mathrm{i}\phi(\vec{h})}$ (same for $\mathbf{F}_{obs}(\vec{h})$ and $\mathbf{F}_{cal}(\vec{h})$),

$$P(F_{obs}(\vec{h}) \mid \mathbf{F}_{cal}(\vec{h})) = \frac{1}{\pi\varepsilon\sigma_\Delta^2} \iint F(\vec{h}) \cdot P_{meas}(F_{obs}(\vec{h}) \mid F(\vec{h})) \cdot P_{phase}(\phi(\vec{h})) \cdot e^{\frac{-[F(\vec{h})e^{\mathrm{i}\phi(\vec{h})} - D\mathbf{F}_{obs}(\vec{h})]^2}{\varepsilon\sigma_\Delta^2}} \mathrm{d}\phi(\vec{h}) \mathrm{d}F(\vec{h})$$

Equation 2-16 Detailed expression of likelihood of one observation given a model

Inserting Equation 2-16 to Equation 2-15 yields the maximum likelihood target energy. According to the data type, there are totally three variants of Equation 2-16.

- MLF target function, for data expressed by diffraction amplitudes[8]
- MLI target function, for data expressed by diffraction intensities[8]
- MLHL target function, for data with experimental phase information[9]

For macromolecular crystallography, Maximum Likelihood target function is superior over Least Squares.

2.2.2. Stereochemistry energy

Structure geometry, e.g., bond length or bond angle of chemical groups in a molecule, often possesses standard values[10]. Explicitly introducing stereochemistry energy to the total target function is indispensable for refinement with low resolution data, as those data are not informative enough for maintaining detailed geometry during the refinement process.

Despite of the rigor of many chemical bonds, angles, dihedrals, etc., stereochemistry energy is typically added in the form of restraint instead of constraints. They are different in definition, computational treatment, and the way data-to-parameter ratio is influenced.

2.2.2.1. Restraints and Constraints

Restraints and constraints both imply a relationship in which a parameter always tends to approach or to be an 'ideal value'. A constraint imposes a 'hard' equality that their difference must strictly vanish at all times. This introduces dependence between associated refinable parameters and equivalently decreases the number of independent refinement parameters. Suppose the number of experiment data, refinable parameters and constraints N_D, N_P and N_C, respectively. The data-to-parameter ratio γ is increased by considering constraints.

$$\gamma_{original} = \frac{N_D}{N_P}$$

Equation 2-17 Data-to-parameter ratio

$$\gamma_C = \frac{N_D}{N_P - N_C}$$

Equation 2-18 Data-to-parameter-ratio with constraints

A restraint, on the other hand, is 'soft' and usually takes the expression of a harmonic energy of which the ideal value is the equilibrium value, and the parameter is allowed to deviate from equilibrium, incurring a restoring force that drags it back. The degree of softness depends on the potential coefficient and will decrease with increasing coefficient. These restraints provide additional information between independent refinement parameters and equivalently serve as extra data.

$$\gamma_R = \frac{N_D + N_R}{N_P}$$

Equation 2-19 Data-to-parameter ratio with restraints

2.2.2.2. Restraint for bond lengths

Atom pair interacting with each other via a chemical bond should be close to the standard value of that chemical bond type. Total stereochemistry bond energy should be the summation of all bond length restraints. w_{bond} is associated weight assigned to each term. A similar weight is present for other restraint classes.

$$E_{bond} = \sum_{bonds} w_{bond}(d_{model} - d_{ideal})^2$$

Equation 2-20 Bond length restraint energy

2.2.2.3. Restraint for bond angles

Similar to Equation 2-20, bond angle restraint is defined as

$$E_{angle} = \sum_{angles} w_{angle}(\omega_{model} - \omega_{ideal})^2$$

Equation 2-21 Bond angle restraint energy

2.2.2.4. Restraint for dihedral angles

Dihedral angle is the angle between two planes defined by four atoms. Suppose that atoms are sequentially labeled as A, B, C, D. Dihedral angle θ is the angle between plane ABC and plane BCD.

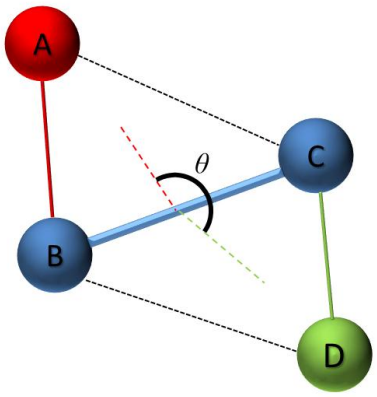


Figure 2-3 Dihedral angle of four sequential atoms

The dihedral angle energy is

$$E_{dihedral} = \sum_{dihedrals} w_{dihedral}(\theta_{model} - \theta_{ideal})^2$$

Equation 2-22 Dihedral angle energy restraint

2.2.2.5. Restraint for planarity

Planarity can be maintained by defining related improper angles (a dihedral angle with torsion axis not a chemical bond) and fix them to 0 or 180 degree. This could be inefficient when many atoms are involved. Another planarity restraint energy is defined to penalize out of plane conformation atoms[11].

$$E_{planarity} = \sum_{groups} w_{plane} \sum_{i} g_i^2$$

Equation 2-23 Planarity restraint energy

g_i is the orthogonal distance of i th atom within group from the plane defined by all atoms within this group via least squares. Double summations are taken over all atoms within a group and all planar groups.

2.2.2.6. Restraint for chirality

An asymmetric carbon atom leads to chirality in a molecule which is non-superposable with its mirror image. A quantity called chiral volume is defined to describe the chirality of an atom (e.g. C_α).

$$V_{model} = (\overrightarrow{r_N} - \overrightarrow{r_{C_\alpha}}) \cdot \left[(\overrightarrow{r_C} - \overrightarrow{r_{C_\alpha}}) \times (\overrightarrow{r_{C_\beta}} - \overrightarrow{r_{C_\alpha}}) \right]$$

Equation 2-24 Chiral Volume of a C alpha atom

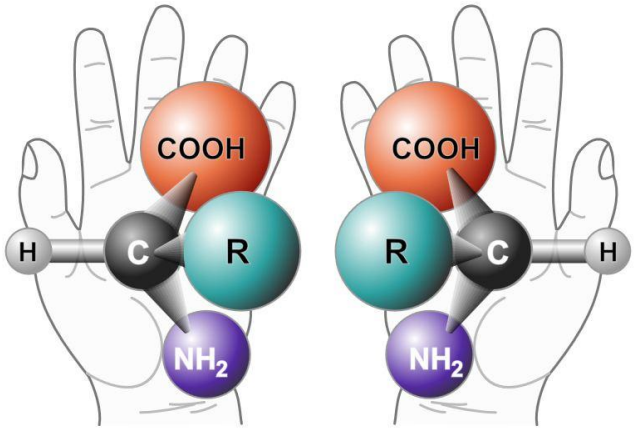


Figure 2-4 Illustration of chirality with C alpha atom of an amino acid molecule[1]

with the chirality energy restraint

$$E_{chiral} = \sum_{chiral} w_{chiral}(V_{model} - V_{ideal})^2$$

Equation 2-25 Chirality energy restraint

2.2.2.7. Non-bonded restraint energy

Van der Waals interaction and electrostatics interaction are merged into a single non-bonded energy term

$$E_{non-bonded} = \sum_{nonbonded-pair} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} + \frac{Cq_1q_2}{r_{ij}} \right)$$

Equation 2-26 non-bonded restraint energy

The Van der Waals energy accounts for both a repulsive term $\frac{A}{r_{ij}^{12}}$ and an attractive term $-\frac{B}{r_{ij}^{6}}$.

2.2.3. *a priori* knowledge based restraints

A priori knowledge can be borrowed to restrain appropriate properties of a model structure, especially for refinements at low resolution, for the purpose of ensure important features (e.g. the secondary structure) that are otherwise difficult to reveal simply from the experiment data.

- Reference model. By analyzing a high resolution homologous model, select features can be analogous and used for refining the target model with low resolution data. An example is the DEN refinement[12].

- Secondary structure restraints. H-bond restraints are imposed to maintain alpha helices, beta sheets and DNA/RNA base pairs.

- Ramachandran restraints. Steer outliers in a Ramachandran plot[13] graph to a favored region to fix incorrect dihedral angle pairs (φ,ψ).

- NCS restraints. Non-crystallographic symmetric copies of a molecule should ideally have identical structures. Geometry difference between each copy can be restrained.

2.3. Refinement target optimization

Refinement target is an energy potential with complicated landscape, numerous local minima and maxima across the conformation space. By optimizing this target, we are essentially trying to explore the global minimum of this unknown landscape.

2.3.1. Gradient minimization

Gradient minimization is a method that by calculation of a set of 'effective force' from each atom's local gradient of potential environment, the conformation is driven under this force and thus moving along the downhill of the landscape until arriving at the closest local minimum. Force on the i th atom is

$$\overrightarrow{F_i} = -\nabla_i E(\vec{r_1}, \vec{r_2}, \cdots, \overrightarrow{r_i} \cdots)$$

Equation 2-27 Force derived from the gradient

This minimization strategy reveals the nearest path towards a local minimum, however, it is unable to overcome an energy barrier, has no access to the global minimum and therefore not used for minimizing a sophisticated target potential.

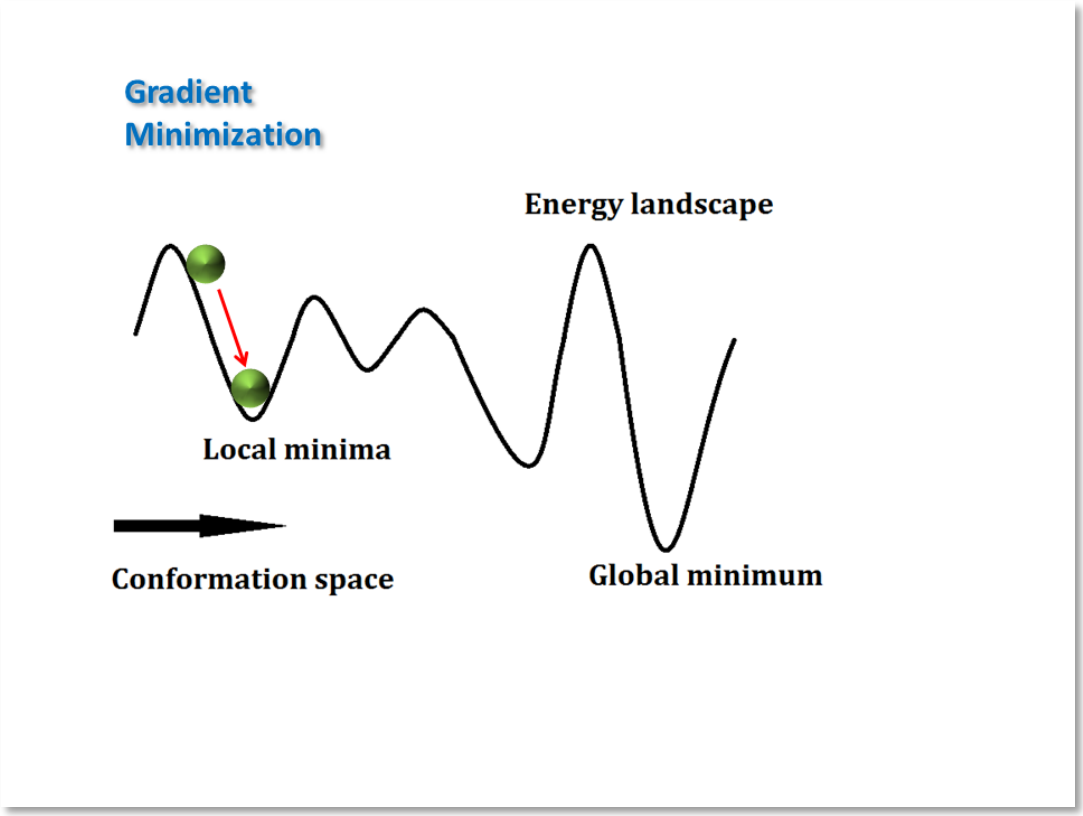


Figure 2-5 Optimization of a target function with sophisticated energy landscape using gradient minimization

2.3.2. Simulated Annealing

Essence of simulated annealing optimization is revealed by the name: a simulation to the annealing process. The latter is known as first melting a solid to liquid phase, followed by a long process of gradually cooling so that all particles are arranged in the lowest energy state.

In computation implementation, simulated annealing is usually controlled by two temperature parameters – the starting temperature and the cooling rate (assuming that we are decreasing the temperature with a constant rate). Generally a high enough initial

temperature would assign particles with large velocities that are necessary to climb over high barriers, but may also lead to system 'blowout'. On the other hand, a quite slow cooling process would make the conformation search for global minimum finer, but may dramatically increase CPU time.

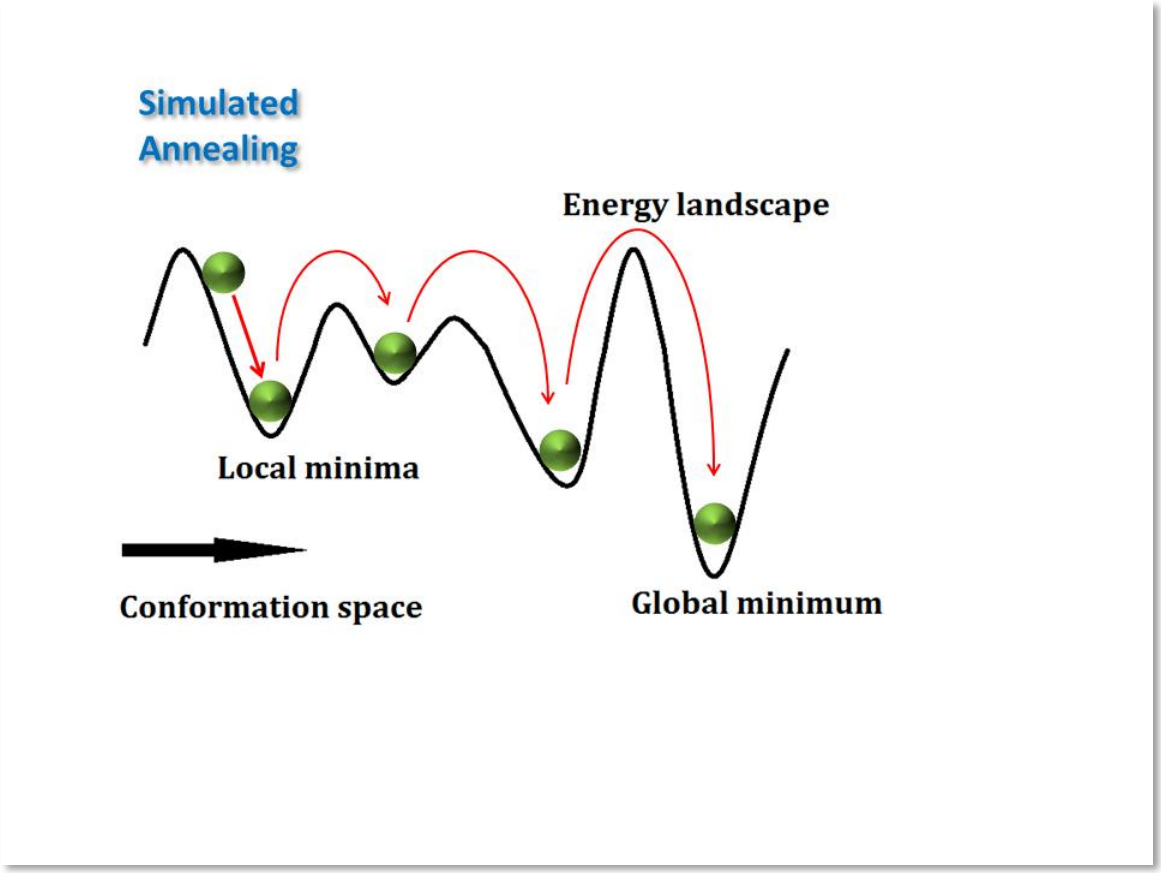


Figure 2-6 Optimization of a target function with sophisticated energy landscape using simulated annealing

2.3.3. Grid search in conformation space

As expected, radius of convergence for grid search should be the largest as the high energy barriers make no difference to any other point of the landscape, because of

the fact that the conformation change is realized directly from switching parameters among different conformation space grid points within a pre-defined sample range. The obvious limitation is that this method is computationally intractable and impractical for systems with large number of parameters (N-dimensional search), or systems with parameters that need a very broad sample range (search space with large dimensions).

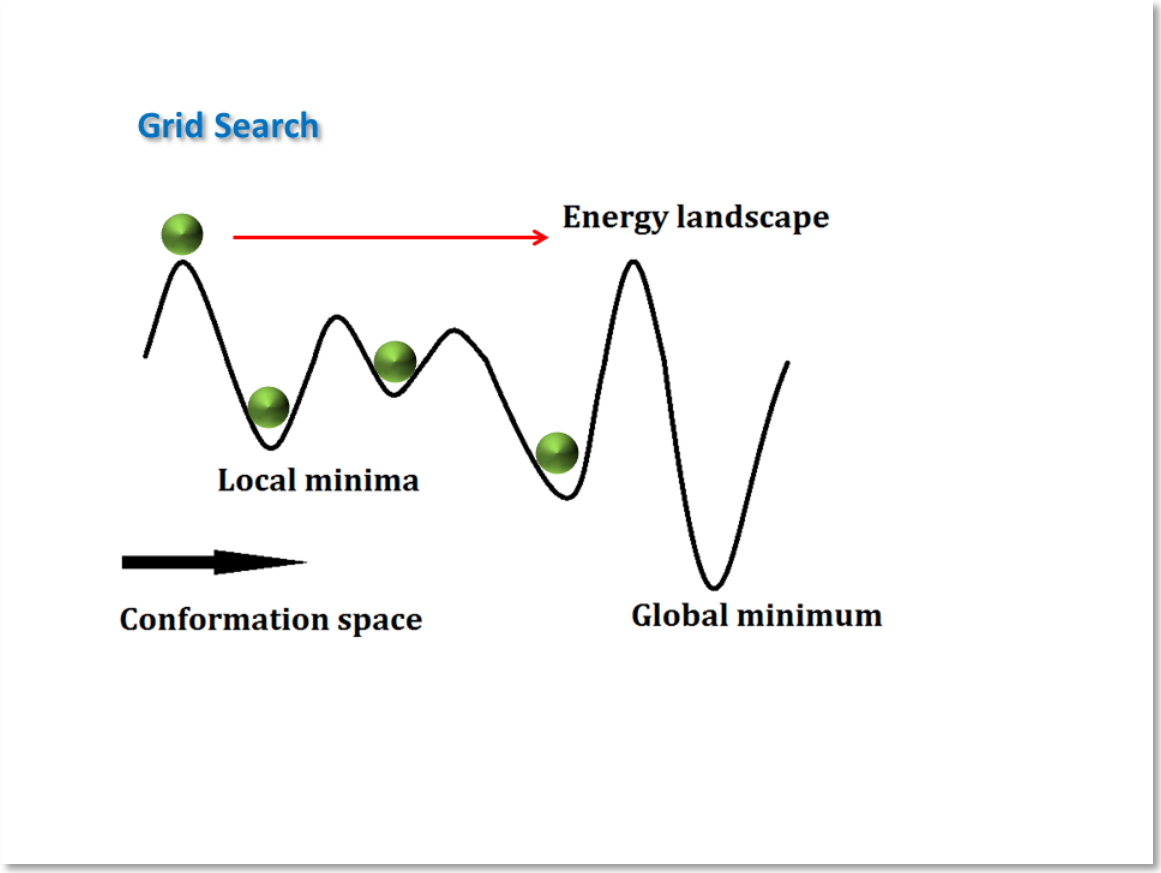


Figure 2-7 Optimization of a target function with sophisticated energy landscape using grid search

As an application with less than three parameters, bulk solvent correction mask model uses the grid search technique to determine a parameter pair k_{sol} and B_{sol} .

Similarly, the Deformable Complex Network approach uses a 3D grid search to determine a DCN parameter configuration that delivers the lowest R free value[14].

2.4. Refinement progress indicators and validation tools

2.4.1. The R value and over-fitting problem

2.4.1.1. The R value

In crystallography, R factor is the common quantity defined to evaluate progress of refinement and quality of a structure[15].

$$R = \frac{\sum_{\vec{h}} \left| F_{obs}(\vec{h}) - kF_{cal}(\vec{h}) \right|}{\sum_{\vec{h}} F_{obs}(\vec{h})}$$

Equation 2-28 Definition of R value

Here k is a scaling coefficient. Smaller R factor indicates a better agreement between observed and calculated amplitude profiles, thus improves the structure for better explanation of experiment results.

2.4.1.2. Over-fitting problem and R_{work}, R_{free}

For macromolecular refinements, number of refinable parameters may exceed that of the diffraction entries. When this happens, a structure model can be ‘over-fit’ during an intensive optimization of the parameters.

The root cause of over-fitting is the undetermined nature of the equation group when independent parameters are more than applicable conditions. Thus, the R value can be made arbitrarily small for a refinement with poor observation-to-parameter ratio.

Cosmetic decrease in R value does not necessarily means an improvement in structure. Artificiality in R value invalidates its objectiveness and makes it unsuitable as a refinement indicator. To address this issue, the idea of cross validation from statistics is introduced[14] for assessment of structure quality.

For this purpose, the entire diffraction data are divided into two sets. One is the working set, which takes up 90%-95% of the data and actually serves as the experiment data used in refinement. The other is the free set generated by a random selection of 5%-10% diffractions from the data pool. This set does not participate in the refinement process and is usually recorded by a 'free flag' for identification in an experiment data file. Therefore, summations over entire data set in all reciprocal space target functions previously discussed are now actually carried out only over the working subset data.

Accordingly, two variants of Equation 2-28, labeled as R_{work} and R_{free} are defined by simply doing the summation over respective data sets.

$$R_{work} = \frac{\sum_{\vec{h}\in \text{working set}} \left| F_{obs}(\vec{h}) - kF_{cal}(\vec{h}) \right|}{\sum_{\vec{h}\in \text{working set}} F_{obs}(\vec{h})}$$

$$R_{free} = \frac{\sum_{\vec{h}\in \text{free set}} \left| F_{obs}(\vec{h}) - kF_{cal}(\vec{h}) \right|}{\sum_{\vec{h}\in \text{free set}} F_{obs}(\vec{h})}$$

Equation 2-29 Definition of the work and free R values

High correlation is expected and observed between the Least Squares target function in Equation 2-13 and R_{work} in Equation 2-29. As refinement proceeds, Least Square target function is continuously being minimized while R_{work} drops down at the same time. Similar correlation exists between the Maximum Likelihood target and the R_{work} as well. R_{free}, on the other hand, is not biased by the model or the refinement procedure, and used as a primary indicator of structure improvement in modern refinements.

2.4.2. Root Mean Square Deviation (RMSD)

RMSD between two related structures is defined to quantitatively assess the overall difference between two coordinates set, or the 'deviation' of one coordinate set from the other, usually after a global alignment.

Suppose two molecule structures with identical number of atoms. The RMSD[16,17] for equivalent atoms is

$$RMSD(1,2) = \sqrt{\frac{\sum_{i=1}^{n} (\overrightarrow{r_{1i}} - \overrightarrow{r_{2i}})^2}{n}}$$

Equation 2-30 Definition of RMSD

In cases when $(\overrightarrow{r_{11}}, \overrightarrow{r_{12}}, ..., \overrightarrow{r_{1i}}, \ldots, \overrightarrow{r_{1n}})$ and $(\overrightarrow{r_{21}}, \overrightarrow{r_{22}}, \ldots, \overrightarrow{r_{2i}}, \ldots, \overrightarrow{r_{2n}})$ denotes structures of only main chain atoms instead of all atoms, the quantity is referred to main chain (or backbone) RMSD.

2.4.3. Global Distance Test (GDT) score

The disadvantage of RMSD roots from the equality of weights for all atom pairs. Certain regions of a molecule, for instance, the loop region, may be quite flexible with various allowed and favorable conformations. Deviation between different conformations can be significant and make considerable contribution to the RMSD, while the structure may in essence be pretty good compared to the impression a hefty RMSD gives.

To address this issue, a GDT score[18] is calculated as the proportion *P* of equivalent alpha carbons in two structures with distances smaller than a defined cutoff (1Å, 2Å, 4Å, 8Å).

$$GDT\_TS = \frac{P(<1\text{Å}) + P(<2\text{Å}) + P(<4\text{Å}) + P(<8\text{Å})}{4}$$

Equation 2-31 Definition GDT Total_Score

Distance cutoff is usually directly labeled with the score name. For example, the GDT(<1Å) score used in the Deformable Complex Network approach below is a GDT score with a cutoff of 1Å.

$$GDT(<1\mathring{A}) = P(<1\mathring{A})$$

Generally, GDT score increases with increased cutoff radius.

2.4.4. TMscore

Zhang *et.al.* proposed a template/model score (TMscore)[19] to eliminate the system size dependence of GDT scores for random structure pairs. GDT scores of 3656 protein pairs with sequence identity less than 30% were computed and a power law between GDT score and protein length was observed.

TMscore, another quantity measuring structure similarity between two proteins, is a variant of the Levitt-Gerstein score[20],

$$TMscore = Max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

Equation 2-32 Definition of TMscore

with L_N and L_T the length of native structure and aligned residues to the template, respectively. d_i is the distance of i th pair of aligned residues. d_0 is a normalized term,

which takes the following form rather than a constant as in other approaches[20-22] in order to eliminate protein size dependence.

$$d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$$

Equation 2-33 d_0 as a function of L_N in TMscore

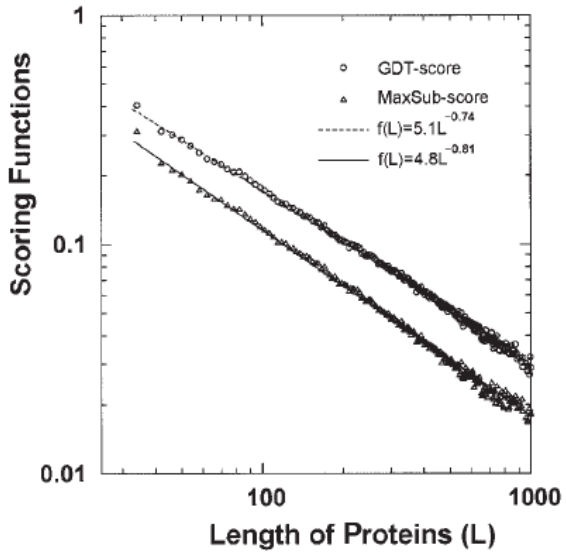


Figure 2-8 Relationship between GDT (and MaxSub) score for random structure pairs and length of proteins[19]

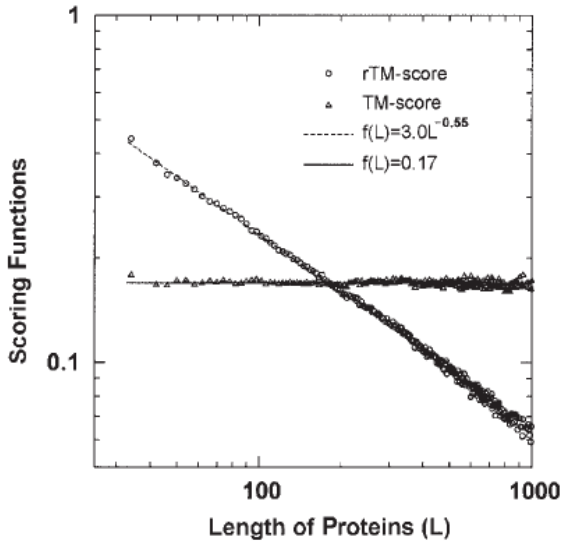


Figure 2-9 Relationship between TMscore (and rTMscore with $d_0 = constant$) for random structure pairs and length of proteins[19]

2.4.5. Ramachandran Statistics

Ramachandran Statistics calculates the percentage of dihedral angle pairs (ϕ,ψ) falling in the favorable regions according to Ramachandran plot[13]. With the definition of dihedral angles discussed in 2.2.2.4, ϕ is the dihedral angle spanned by two planes formed by atom $C_{i-1} - N - C^{\alpha}$ and $N - C^{\alpha} - C_i$. Similarly, for ψ it is spanned by $N_i - C^{\alpha} - C_i$ and $C^{\alpha} - C_i - N_{i+1}$. Pairs for all residues are plotted on a two-dimensional map and pairs falling within several favored regions (determined by the characteristics of alpha helices and beta sheets) of the map are counted. The percentage is defined as the Ramachandran Statistics and is used as an indicator of secondary structure quality.

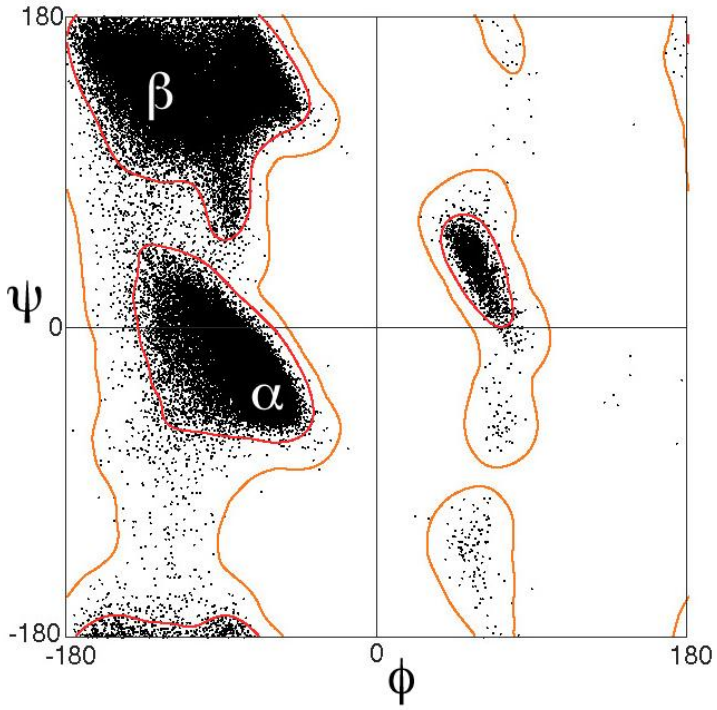


Figure 2-10 An example of Ramachandran Plot[1]

As shown in Figure 2-10, black dots represent (ϕ,ψ) pairs, red circles sketch the favored region, while orange lines sketch the allowed region. Ramachandran Statistics is defined as

$$Ramachandran\ Statistics = \frac{N_{\text{pairs in favored regions}}}{N_{\text{total pairs}}}$$

Equation 2-34 Calculation of Ramachandran Statistics

The higher the Ramachandran Statistics is, the better the secondary structure of a structure model is expected.

2.4.6. Electron Density Map

Electron density map is calculated based on Equation 1-13

$$\rho(\vec{r}) = \frac{1}{(2\pi)^3} \int_{diffractions} \mathbf{S}(\vec{h}) \cdot e^{-i\vec{h}\cdot\vec{r}} \, d\vec{h} = \frac{1}{(2\pi)^3} \int_{diffractions} S(\vec{h}) e^{i\theta} \cdot e^{-i\vec{h}\cdot\vec{r}} \, d\vec{h}$$

Several density maps can be obtained depend on what data are used for substituting

- F_C map – Electron density calculated with model structure factor and phase, which is a map solely related to the current model
- F_O map -- Electron density calculated with experiment amplitude and model phase, which shows the observed electron density
- F_O-F_C—Difference between the observed and model density map, which tends to be zero when model is correct, moderate non-zero when incorrect atom type is modeled, large positive when an atom is missing from the model and large negative when model contains an atom but supposedly not.
- $2F_O$-F_C—Summation of observed and difference density map, widely used for model building and structure validation.

Chapter 3

Deformable Complex Network Approach

3.1. Motivation and a brief summary

It is often a challenge to atomically determine the structure of large macromolecular assemblies, even if successfully crystallized, due to their weak diffraction of X-rays. Effective number of diffractions available for structure determination looks small, especially when we search for the optimum conformation in the conventional coordinate space. It is required to reduce the number of degrees of freedom thus make the observations/variables ratio greater than one for the theoretical solvability of biomolecular crystallography. Therefore, either a torsion-angle based molecular dynamics method[23] that specifies the torsion angles of a protein as the degrees of freedom, or a normal mode analysis[24] which describes the motion of proteins using a small set of low frequency normal modes at a fairly acceptable accuracy, can be chosen as an ideal tool for the crystallography with low resolution data. Moreover, interpretation of the experiment data to predict the structure is often hindered by the limited agreement

between them. There has always been a need for low resolution structures to be determined at higher accuracies in order to allow valid and intensive researches on the function of those molecules.

This work combines the torsion-angle protocol with the deformable complex network (referred to as DCN) approach, to further derive and make use of useful information from a pre-determined homologous or comparative protein model. It can be shown that, by merging the information independently fetched from the deformable angular network (DAN) and the DEN[12,25] thus generating a DCN model, there is still room for additional improvement to macromolecular refinements over the existing DEN method[12], by a boost from 13% to 264% as assessed by the free R value[14]. Firstly, in order to objectively evaluate the quality of the refined structure, we performed a full refinement against an experiment data set (without experimental phase information) that already has a high resolution (1.8Å) structure deposited into the Protein Data Bank. The data set was then truncated to three different limits to synthesize three lower resolution data sets. We used those data sets for subsequent refinement and compared the results with the existing high-resolution structure (‘true structure’). Improvements are observed across multiple criteria, from the R_{free} value, to the all-atom Root Mean Square Deviation(RMSD), the GDT (<1Å) score[18] and the TM score[19]. Further improvement is expected with the availability of the non-crystallographic symmetry (NCS) information, as well as phase information from experimental methods such as heavy atom isomorphous replacement[26]. Secondly, to ensure generality, we also randomly selected sixteen low resolution structures from the Protein Data Bank and performed re-refinements with those deposited experiment data. Consistent improvements by DCN

have been seen over conventional refinement and standalone DEN refinement, as indicated by the R_{free} value, the Ramachandran statistics[27] as well as the calculated phase combined electron density map.

3.2. Method

3.2.1. Summary

Starting from a given protein's sequence (target sequence) information, we first individually performed a FASTA search to each chain of the molecule. Templates that shared higher sequence identity with the target sequence and possessed higher resolution would be preferable. Five homology structure candidates were subsequently constructed to a chosen template and one with the lowest DOPE score was picked and served as a reference model for a certain chain of the molecule. After reference structures for all chains (if applicable) of a molecule had been built, these structures were merged together and served as the only reference model for the whole molecule. DCN excludes all sorts of inter-chain atoms' interactions when deformable angular and elastic network models are defined. As a result, rather than in the target molecule, different chains in the reference 'molecule' are independent and can take whatever relative positions and orientations. The DCN model and corresponding restraints were automatically generated according to a pre-set criteria for angular network triplets and elastic network pairs. These restraints contribute to the E_{DCN} term in the total energy function (target function, see below). Simulated annealing[28] was used as the refinement protocol, with a starting temperature of 3,000K and a cooling rate of 50K per step. The torsion-angle dynamics[23] was performed

as the MD. Refinement with each parameter group was repeated ten times with different random seeds for initial velocities assignments and DCN restraints selections.

3.2.2. Target function

The target function takes the following form

$$E_{target} = E_{stereo} + w_a \cdot E_{experiment} + w_{DCN} E_{DCN} \quad (1)$$

E_{stereo} is the usual stereochemisty energy that regularizes bond lengths, bond angles and others to those pre-defined, well-accepted standard values. $E_{\text{experiment}}$ is the experimental term that makes use of the reflection data, and its weight w_a is determined automatically and adjusted frequently to ensure that the force derived from the experiment term is approximately of the same order of magnitude to that of the sum of other terms in this equation. Typically the amplitude-based maximum likelihood function (MLF) would be used instead of the conventional crystallographic residual (least square). In case phase information is obtained through experimental techniques, a refinement is executed with the MLHL target function where experiment phase contributes in the form of Hendrickson Lattman coefficients[29]. E_{DCN} is the potential energy that roots from the deviation of selected atom pairs and triplets in the target molecule from their corresponding equilibrium values; these values are derived from both the reference models as well as the target's current structure itself.

3.2.3. DAN and DCN approach

3.2.3.1. Introduction

The deformable angular network (DAN) model is a model consists of a series of angles, each spanned by two bonds within an atom triplet which is to be found by referring to the same chain ID, residue number and atom name in both the reference and target structures. Generally, when a reference structure is obtained via homology modeling[30,31] , all of the atoms that have equivalent atoms in the target structure are picked as candidate elements of DAN triplets. Further filtering of the triplets arises from the requirement that, 1) vertex atom of the triplets has interactions with both tail atoms with a user-defined search radius or one equal to the upper distance cutoff of DEN restraints, 2) vertex atom and each tail atom should be no more that ten residues apart. After the first two rounds of preliminary selection, the remaining eligible triplets are subject to a third one based on the vertex angles that spanned by the two "line" connecting the vertex and each tail. We choose those angles that have a value between 60 and 120 in degree. The final angular restraints for later refinement purpose are randomly selected from the triplets pool. This random behavior is determined by a seed pre-defined before refinement. The total number of restraint entries is determined according to a multiplicity factor (typically 1) and the number of atoms that participate in the generation of the DAN model. Since interactions between different chains are excluded, DAN is usually made chain by chain and then merged into a single reference model. The selection of atoms for DAN generation can also be constrained to any groups of atoms or their combinations within the same chain. Deformable complex network (DCN) is established if both DAN and deformable elastic network (DEN) are present. In this case,

two models are built independent, even from different results of homology modeling. However, they work together to lead the direction of conformation search throughout the refinement process. The restraints of DAN and DEN contribute in a reciprocal way to improve the refinement and final structure. These restraints are considered as unified DCN restraints and added to the total refinement target function.

3.2.3.2. DAN/DCN modes

Due to the ways the DAN triplets pool is initially built up, there are two modes for DAN (therefore for DCN as well). The first is the directional mode. After all atoms have been serialized (i.e., the atom serial number is assigned to each atom in PDB), the vertex atom seeks interaction only with atoms of higher atom serial numbers. This mode results in a fact that in each triplet, vertex atom has the lowest atom serial number (Figure 3-1). Hence, the directional mode is only capable of covering a fraction of selections from the entire candidates pool. In the other hand, this mode also comes with a feature that explicitly eliminates an unfavorable situation, in which each of the three atoms within a single triplet is picked as the vertex one after another, and three resultant restraints actually correspond to three interior angles of the same triangle. This kind of restraints is considered too strong, especially when the reference model has a low quality and is not reliable. The second is called the arbitrary mode(Figure 3-2). In this case, direction in which the vertex points to a tail atom is not restricted at all. This allows a hundred percent coverage of all angles for possible selection from the triplets pool. The definition of arbitrary mode ensures that no useful information from reference model is discarded from the beginning, therefore, chances are increased that a list of better atom triplets will be established and serves as the DAN restraints. In a DAN reference file, for

both modes, each restraint entry is listed in the order of ‘vertex atom - first tail atom - second tail atom - angle value’. The atom serial number of the first tail atom is constantly lower than that of the second to prevent restraints duplications. Typically the arbitrary mode is chosen as the default mode. However, it is wise to sometimes perform the DCN refinement in directional mode to achieve a possibly lower R_{free}.

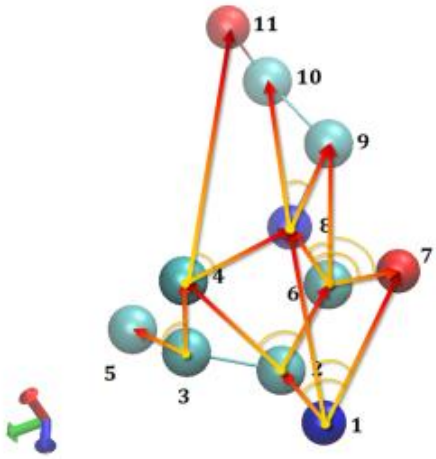


Figure 3-1 Directional mode (D-mode) of DAN/DCN

Figure 3-1： The directional mode (D-mode) of DAN/DCN. Atom serial number of the vertex is always lower than that of the first and second tail atom. For example, for triplet 3-4-5, in directional mode, the only angle that can be selected is ∠ 435, where atom 3 is the vertex. Therefore, no more than one angle will be restrained for a given atom triplet and the directional mode tends to ‘spread’ over the entire structure and lead to more different atoms being included in the final DAN restraint list.

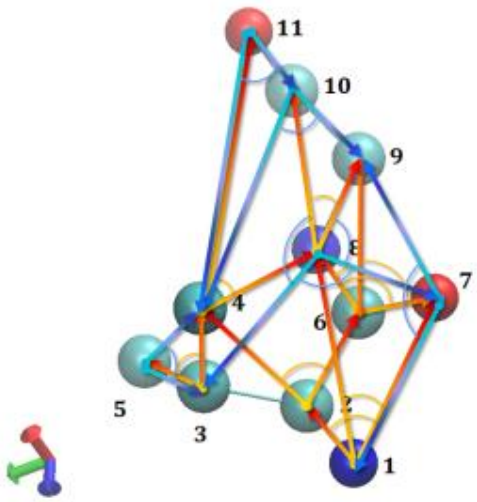


Figure 3-2 Arbitrary mode (A-mode) of DAN/DCN

Figure 3-2： The arbitrary mode (A-mode) of DAN/DCN. No restriction is placed for angle selection when a certain atom triplet has been picked. For triplet 3-4-5, in addition to ∠ 435 that can also be shot by directional mode, arbitrary mode as well allows angles such as ∠ 354 (shown), where atom 5 is the vertex, and ∠ 345 (not shown), where atom 4 is the vertex. The arbitrary mode will include all possible angles present in a structure. If the cutoff criterion for DAN is flexible enough, two or even all three angles within the same triplet may be eligible for candidacy for final restraints selection. As a result, arbitrary mode is possible to target angles that have been excluded by directional mode at the first place, but may create less atom diversities than the restraint list extracted from directional mode DAN pool. The generated DAN restraint file lists the triplet in the order of vertex, first tail and second tail. For both modes, the atom serial number of the first tail is by definition lower than the second to avoid selection duplication.

3.2.3.3. DCN energy restraint equations

The DCN potential is the sum of the harmonic bending energy of DAN and stretching energy of DEN.

$$E_{\mathrm{DCN}} = k \cdot E_{\mathrm{DAN}} + E_{\mathrm{DEN}} \qquad 2(a)$$

$$E_{\mathrm{DAN}} = \sum_{i,j,k} (\theta_{ijk} - \theta^0_{ijk}(\mu, n))^2 \qquad 2(b)$$

$$E_{\mathrm{DEN}} = \sum_{l,m} (d_{lm} - d^0_{lm}(\gamma, n))^2 \qquad 2(c)^{12}$$

The summations are taken over all angle triplets for DAN and all distance pairs for DEN. θ_{ijk} and d_{lm} are the instantaneous angle for an atom triplet and distance for an atom pair at a conformation state during the refinement, respectively. $\theta^0_{ijk}(\mu, n)$ and $d^0_{lm}(\gamma, n)$ are the corresponding equilibrium angle and distance at a specific (n th, see blow) refinement step. We set the coefficient *k* to 0.01 as the angles are in degree. $\theta^0_{ijk}(\mu, n)$ and $d^0_{lm}(\gamma, n)$ are updated every six MD steps (when the temperature also drops 50*K*) according to the following equations,

$$\theta^0_{ijk}(\mu, n+1) = (1-\phi) \cdot \theta^0_{ijk}(\mu, n) + \phi \cdot \left[\mu\theta_{ijk} + (1-\mu)\theta^{ref}_{ijk} \right] \qquad 3(a)$$

$$d^0_{lm}(\gamma, n+1) = (1-\kappa) \cdot d^0_{lm}(\gamma, n) + \kappa \cdot \left[\gamma d_{lm} + (1-\gamma) d^{ref}_{lm} \right] \qquad 3(b)^{12}$$

The angle and distance's next equilibrium values $(\theta^0_{ijk}(\mu, n+1), d^0_{lm}(\gamma, n+1))$ are functions of their current equilibrium values $(\theta^0_{ijk}(\mu, n), d^0_{lm}(\gamma, n))$, their actual instantaneous values (θ_{ijk}, d_{lm}), as well as values of equivalent triplet and pair in the reference model $(\theta^{ref}_{ijk}, d^{ref}_{lm})$. For numeric stability, typically, the initial equilibrium values

of the atom triplet $\theta_{ijk}^{0}(\mu,0)$ and pair $d_{lm}^{0}(\gamma,0)$ are set to be equal to those values in the starting structure. ϕ and κ controls the transition between consecutive equilibrium values. For initial relaxation, ϕ and κ are set to 0 (hence terms including μ and γ vanish) during the first three macrocycles. After that, ϕ and κ are set to a fixed value of 0.1. μ and γ are optimized together with w_{DCN}, the weight of DCN potential, via a 3D grid search. w_{DCN} is set to 0 during the last two cycles to reduce the bias effect of the target energy minimum.

3.2.4. Selection of reference model and (γ, w_{DCN}, μ) parameter group

Having the target protein's primary structure, with *Modeller*[31] one could perform a FASTA search chain by chain against an existing protein database so that several homologous protein chains sharing various sequence identity would be listed out as the templates for the reference model. Typically, candidates with higher identity, higher resolution as well as longer length would be preferable as they generally ensure higher quality. We used the automatic procedure of *Modeller* program throughout the whole modeling process, including the sequence alignment as well as the template building. We chose to build five final models for each template picked, and the one with the lowest DOPE score was designated as the reference model for a chain of the target structure. After all chains or part of them that interest had their reference models built, these models were merged into one PDB file and served as the unique reference model for subsequent use. The parameter group (γ, w_{DCN}, μ) is optimized via a 3D grid search (Figure 3-3) through 180 grid points: (0, 0.2, 0.4, 0.6, 0.8, 1) for γ, (3, 10, 30, 100, 300) for w_{DCN} and

(0, 0.2, 0.4, 0.6, 0.8, 1) for μ. At each point, ten refinements with different random seeds but otherwise identical were carried out and the result with the lowest R_{free} would represent the final refined structure at that grid point. The seed controls the assignment of initial velocities for atoms as well as the selection of DCN restraints from the pair and triplet pool. It should be noted that, due to the relatively stochastic nature of the effect of the refinement, particular practice like using a parameter group with value falling between the closest search grid points, carrying out more refinement repeats (e.g. 20) or simply picking a distinct random integer as the seed(e.g. 18593) had a chance to give considerably better results for select systems (data not shown). However, to ensure consistency, generality and valid comparison, we stuck to the same grip search strategy and used the exact integers from 1 to 10 as the ten random seeds throughout this work.

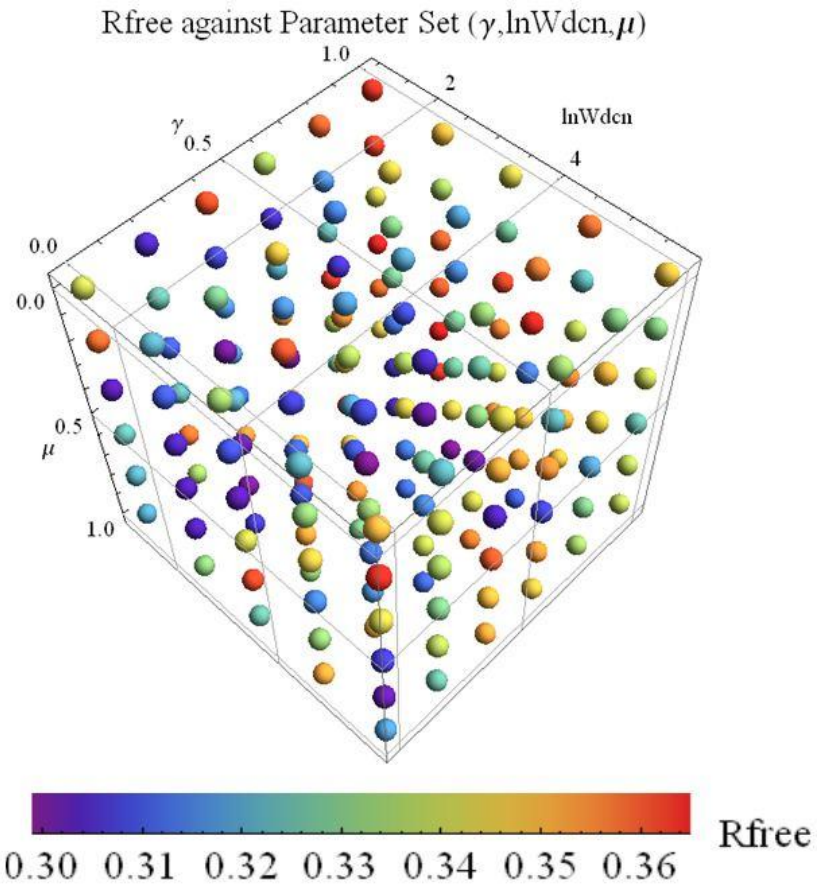


Figure 3-3 3D grid search for best parameter set (γ, w_{DCN}, μ)

3.2.5. Input data preparation before refinement

Many proteins possess non-standard ligands and modified residues, which are usually listed as hetero-atom entries (HETATM) in the PDB files. Currently many of them are not included in the CNS database and a straight refinement with their presence in the initial structure would cause the task to cease. Previous work[12] used an automated import method thus only residues and ligands recognized by CNS were involved. Here, prior to performing the refinement, we fetched the topology and parameter files of those ligands and modified residues from the Hetero-compound Information Center Uppsala

(HIC-Up) and subsequently imported their coordinates for refinement like other compounds of a molecule. In that those ligands diffract X-ray as well, improved agreement with experiment data was expected as manifested by $R_{free.}$ (supplementary Table 3).

For the case of tobacco PR-5d protein (PDB ID 1AUN), the high resolution data set obtained from the Protein Data Bank was truncated using CCP4 software into three lower resolution sets at 3.5 Å, 4.0 Å and 4.5 Å, respectively. These three synthetic sets served as the original experiment data for subsequent refinement and analysis. The starting structure for the refinement could be half-refined, theoretically predicted or manual built, but should be reasonably close to the target structure. In this work, for the full refinement of 1AUN, its reference model (PDB ID 1PCV) was selected as the starting structure, whose positions and orientations were determined by molecular replacement with *Phaser*[32] against each of the three low resolution data sets. Straightforwardly, as for other re-refinement tasks, the starting structure was just the known low resolution structure to be refined. In this work, the lower cutoff value for DEN and DAN selection was set to 3Å and 60°, while the upper 15Å and 120°, respectively. For both DEN and DAN, the sequence separation range was chosen to be 0 to 10 residues, and the restraints/atoms ratio was set to 1. The search probe radius for atom interactions in both DEN and DAN structures was set to be equal to DEN's upper distance cutoff value (15Å) for all cases, except for PDB ID 2VKZ, where the value was set to 13Å in DAN.

3.2.6. Refinement protocol

Torsion angle molecular dynamics[23] (TAMD) with reduced degree of freedom, combined with traditional simulated annealing[28] was used as the main refinement protocol[12]. The time of each MD step was 4*fs*. For the annealing process, the initial temperature was set to 3,000*K*, with a decreasing rate of 50*K* per 6 TAMD steps. Every 6 TAMD steps could be defined as a 'microcycle', which determined the adjustment frequency of both the temperature and each DCN restraint's equilibrium. The period the temperature dropped from 3,000*K* to 0*K* consisted of a 'macrocycle'. Each refinement task in this work, including conventional refinement, DEN refinement and DCN refinement used eight such macrocycles. During the first three of them, φ and κ were set to zero rather than 0.1 to allow initial relaxation. The van der Waals radii had been shrunk to 75% of the original value during several initial macrocycles, together with a reduced van der Waals force constant to facilitate sampling, and were thereafter fully restored in the last two cycles. Moreover, DCN restraint weight was also set to zero at the last two macrocycles to reduce the bias of the target's global minimum. Anisotropic overall B-factor correction and bulk solvent correction were applied for all refinements and no positional minimization used. For the sixteen re-refinement tasks, 50 steps of group B-factor minimization with a ten-fold increase for target sigma values of B-factor main/side chain bonds/angles restraints were performed and the initial values of B factors were reset to 50$Å^2$. Ligands not by default recognized by CNS were explicitly defined as groups for group B-factor minimization. As what had been done with DCN parameters, for appropriate comparison purpose, all these refinement parameter settings were also kept identical across all test systems in this work, even though a different value of a

parameter, for instance, the multiplicity of target sigma value for group B factor minimization (data not shown), the initial temperature or the cooling rate (as one of the most important parameters whenever simulated annealing protocol is introduced), would no doubt be possible to be further optimized for a lower R_{free}. Upon completion of a refinement, all refined structures were sorted according to the R_{free} and one with the lowest value was then picked for subsequent analysis, remodeling or other purposes.

3.2.7. Coding and program

Algorithms of the DCN approach, packed into several source code files and header files, were written fully compatible with version 1.3 of the Crystallography and NMR System (CNS)[33,34] and compiled with Intel Fortran v10.1.051 in this work. The computation was carried out on the Shared University Grid at Rice (SUG@R) cluster platform of the Shared Computing Resources (ShareCoRe). Each refinement task was done on a single core of an Intel Xeon processor running at 2.83GHz. VMD[35]and Coot[36] were used for drawing purpose. TMscore[19] program was used for calculations of GDT(<1Å) score and TMscore. Molprobity[27] was used to evaluate Ramachandran statistics.

3.3. Results and Analysis

3.3.1. Automatic full refinement

In our test case, we used the crystal structure of tobacco PR-5d protein (PDB ID 1AUN) and its experimental data truncated to 3.5Å, 4.0Å and 4.5Å. To allow proper assessment of the DCN approach, we repeated the refinement under exactly the same

protocol, but with two different target functions. One of them is the conventional target, which only combines the stereochemistry potential[10] and the experiment data (in the form of Maximum-likelihood energy[7]). The other is the conventional target plus the DEN potential. Thanks to the presence of the 'true structure' (to the reliability of 1.8Å), in addition to the R_{free}, which measures the fit of the structure to the experiment, we were able to assess the quality of the refined structure by showing the all-atom RMSD from the true structure, the global distance test (GDT) score, and the TM-score. It is shown that, among all three approaches, DCN delivers the most favorable GDT (Figure 3-6) and TM scores (Figure 3-7) among all three refinement approaches, and more accurate (i.e. lower RMSD) structure coordinates than DEN (Figure 3-5). The DCN R_{free} is also significantly improved over DEN and conventional refinement, except at 4.5Å, where DCN has a slightly ($\sim 2 \times 10^{-3}$) higher R-free value than DEN (Figure 1-1). It is noticed that, even in these cases where the R_{free} improvement by DCN is not very remarkable, the actual quality of the DCN refined structure is better than that of DEN (Figure 3-5, Figure 3-6, Figure 3-7). It is noted that at 3.5Å, conventional refinement has the lowest RMSD compared with DCN and DEN. Therefore, DCN is expected to have the best performance for refinement tasks with X-ray resolution data lower than 4Å (Table 1).

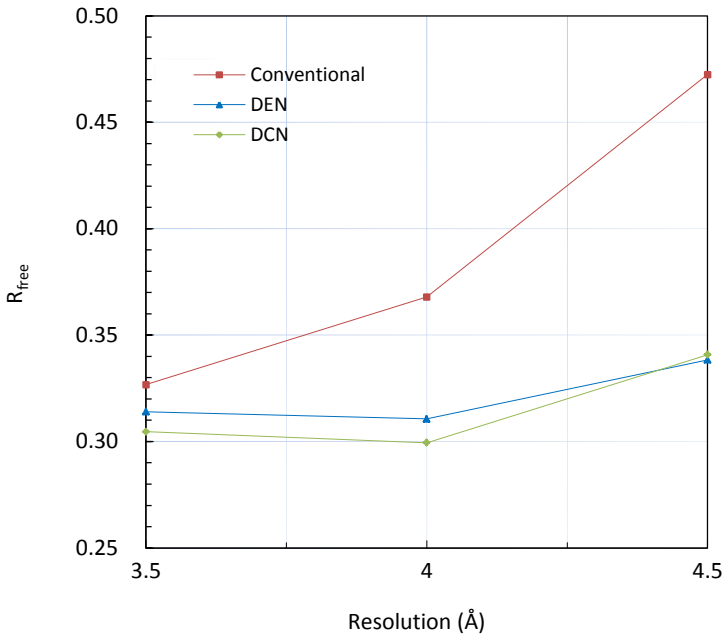


Figure 3-4 R_{free} vs Resolution for Conventional, DEN and DCN

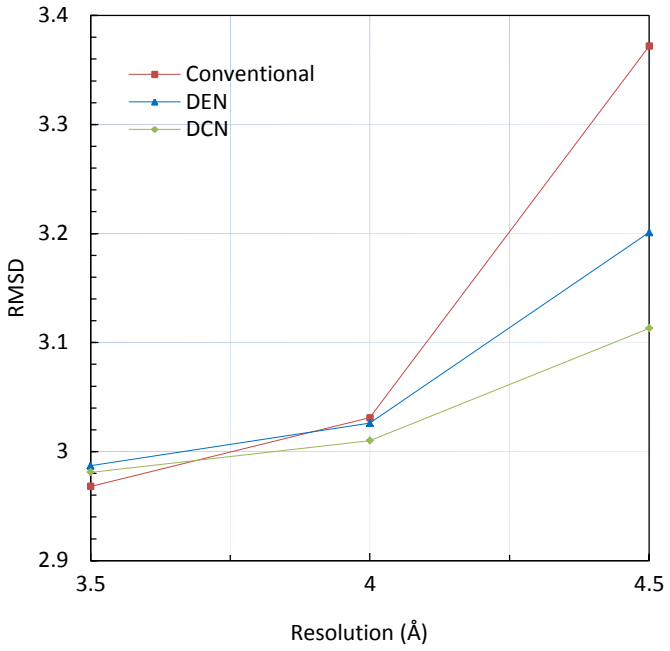


Figure 3-5 RMSD vs Resolution for Conventional, DEN and DCN

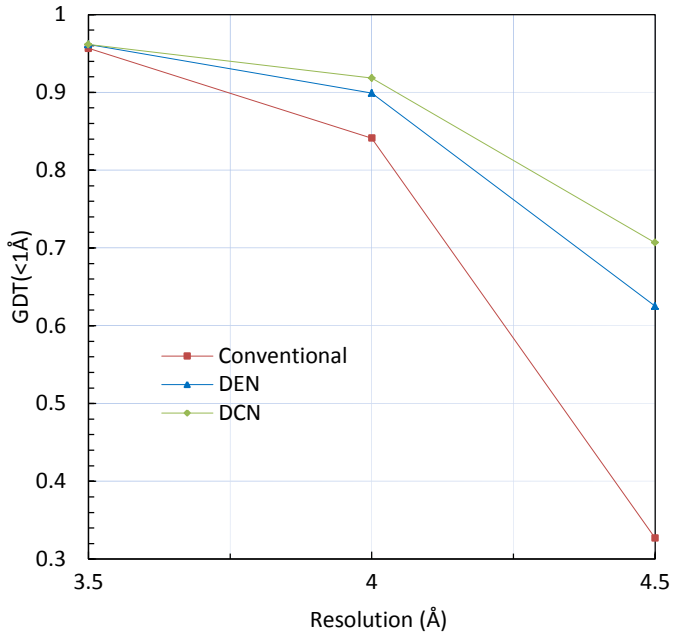


Figure 3-6 GDT(<1Å) vs Resolution for Conventional, DEN and DCN

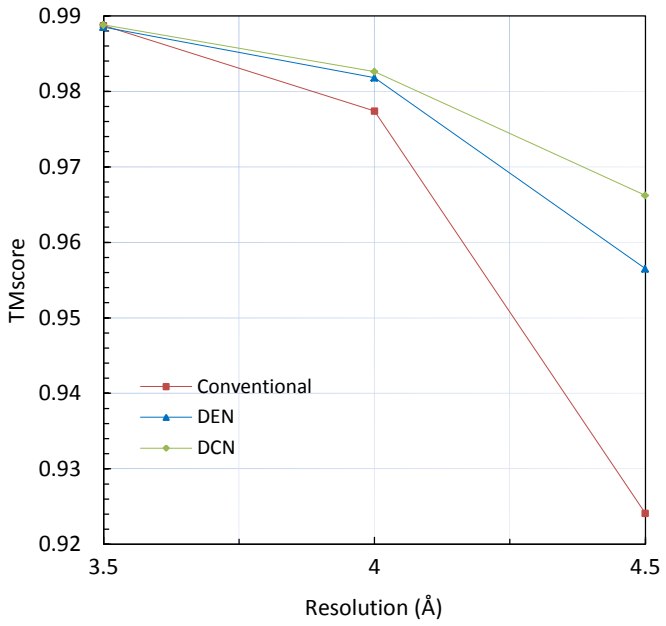


Figure 3-7 TMscore vs Resolution for Conventional, DEN and DCN

Resolution (Å)	Refinement Approach	R_{free}	All-atom RMSD	GDT(<1Å) score	TMscore
3.5	Conventional	*0.3267*	**2.968**	*0.9567*	0.9887
	DEN	0.3140	*2.987*	**0.9615**	*0.9885*
	DCN	**0.3046**	2.981	**0.9615**	**0.9888**
4	Conventional	*0.3679*	*3.031*	*0.8413*	*0.9774*
	DEN	0.3106	3.026	0.8990	0.9818
	DCN	**0.2994**	**3.010**	**0.9183**	**0.9826**
4.5	Conventional	*0.4724*	*3.372*	*0.3269*	*0.9241*
	DEN	**0.3384**	3.201	0.6250	0.9565
	DCN	0.3408	**3.113**	**0.7067**	**0.9662**

Table 3-1 Refinement of tobacco PR-5d protein (PDB ID 1AUN) based on a homology model of a plat antifungal protein osmotin (PDB ID 1PCV) with a sequence identity of (79.51%) and an initial all-atom RMSD of 3.156A to the 'true structure' of 1AUN.

Table 3-1: The conventional, DEN and DCN refinements were carried out at three resolution limits, truncated from the high resolution experiment data set. The best refined structures with lowest R_{free} value by each approach were subsequently subject to three additional validations as all-atom RMSD, GDT(<1Å) score and TMscore to further assess the quality of the structures after each refinement. For each of the total twelve controls (four kinds of score × three resolutions), most favorable results, i.e., lowest R_{free} and all-atom RMSD, and highest GDT (<1Å) and TMscore, are highlighted with bold font, whereas least with italic font. DCN gives ten out of twelve best results, and no worst, while DEN delivers two best results (one of them shared with DCN) but also two worst. Conventional refinement concedes all other ten worst values with only one best (RMSD) at the high resolution.

3.3.2. Automatic re-refinements

We also randomly selected sixteen low-resolution structures (4.0Å - 4.51Å) from the Protein Data Bank and performed re-refinements with the aid of respective high-resolution homology models. All the structures are required to have an all-atom (that is, including side chain atoms) coordinates. For certain structures, the topology and parameter files of non-standard ligands, ions, and modified residues, which are indispensable for refinement to be executed, were obtained from the Hetero-compound Information Center – Uppsala (HIC-Up). To test the performance of the DCN approach, we automatically carried out the re-refinements without any manual inspections, interruptions or manipulations throughout the refinement. In order to minimize bias, we reset the DCN potential to zero at the last two of totally eight refinement macro-cycles.

As a control, identical protocol and settings were used for DEN and conventional refinements, for each of the sixteen re-refinement tasks. These re-refinements enabled a wider and more general comparison between all three methods across the PDB database.

3.3.2.1. Results overview

PDB ID	Resolution (Å)	R_{free}			DCN improvement		R_{free}-R_{work}			Ramachandran Statistics		
		Conventional	DEN	DCN	ΔR_{free} over Conventional	net gain fraction over DEN improvement*	Conventional	DEN	DCN	Conventional	DEN	DCN
1ISR	4.00	**0.2237**	**0.2164**	**0.2110**	0.0127	74%	0.066	0.064	0.061	**0.833**	**0.863**	0.878
1JL4	4.30	0.3700	0.3639	0.3525	0.0175	187%	0.109	0.115	0.111	0.567	0.712	0.718
1R5U	4.50	0.3165	0.3048	0.2983	0.0182	56%	0.056	0.052	0.048	0.646	0.730	0.748
1XXI	4.10	0.3821	0.3224	0.3146	0.0675	*13%*	0.112	0.099	0.094	0.631	0.806	0.800
1YE1	4.50	0.3377	0.3024	0.2936	0.0441	25%	*0.138*	*0.131*	*0.125*	0.781	0.853	**0.905**
1YM7	4.50	0.2764	0.2739	0.2723	0.0041	64%	0.030	0.034	0.033	0.703	0.781	0.751
2A62	4.50	0.3622	0.3548	0.3353	0.0269	**264%**	0.096	0.086	0.069	0.568	0.651	0.670
2BF1	4.00	*0.4866*	*0.4431*	*0.4266*	0.0600	38%	0.086	0.050	0.040	*0.383*	*0.453*	*0.523*
2I37	4.15	0.3646	0.3320	0.3257	0.0389	19%	0.037	**0.012**	**0.003**	0.737	0.851	0.848
2Q7N	4.00	0.2649	0.2621	0.2606	0.0043	54%	0.021	0.019	0.018	0.774	0.768	0.770
2QAG	4.00	0.4052	0.3881	0.3852	0.0200	17%	0.030	0.022	0.020	0.483	0.551	0.573
2VKZ	4.00	0.3117	0.2988	0.2964	0.0153	19%	0.081	0.081	0.080	0.723	0.822	0.830
2YHJ	4.00	0.3734	0.3573	0.3442	0.0292	81%	0.095	0.086	0.082	0.728	0.746	0.836
3ALZ	4.51	0.2501	0.2461	0.2367	0.0134	235%	**0.019**	0.016	0.009	0.667	0.712	0.721
3FUS	4.00	0.4187	0.4057	0.4007	0.0180	38%	0.059	0.044	0.039	0.537	0.563	0.576
3US2	4.20	0.4597	0.4311	0.4239	0.0358	25%	0.129	0.109	0.098	0.399	0.543	0.555
Average	4.20	0.3502	0.3314	0.3236	0.0266	76%	0.073	0.064	0.058	0.635	0.713	0.731
Minimum	4.00	**0.2237**	**0.2164**	**0.2110**	*0.0041*	*13%*	**0.019**	**0.012**	**0.003**	*0.383*	*0.453*	*0.523*
Maximum	4.51	*0.4866*	*0.4431*	*0.4266*	**0.0675**	**264%**	*0.138*	*0.131*	*0.125*	**0.833**	**0.863**	**0.905**

* Net gain fraction over DEN improvement is defined as $\left(\frac{\left|R_{free}^{DCN} - R_{free}^{Conventional}\right|}{\left|R_{free}^{DEN} - R_{free}^{Conventional}\right|} - 1\right) \times 100\%$

Table 3-2 Results of sixteen low-resolution re-refinement tasks.

Table 3-2: R_{free} and its improvement, R_{free}-R_{work} as well as Ramachandran Statistics are shown. Properties of structure, experiment data and reference model of each test system are listed in Supplementary Table 1 and Supplementary Table 2. Out of a total of sixteen test systems, DCN outperforms DEN in sixteen (100%) in R_{free}, sixteen (100%) in R_{free}-R_{work}, and fourteen (87.5%) in Ramachandran statistics. When compared with conventional, these ratios come to 100%, 87.5% and 93.75%,

respectively. Moreover, 87.5% cases achieve an R_{free} improvement over 0.010 by DCN with the largest one of 0.0675 (PDB ID 1XXI).

3.3.2.2. Decrease in R_{free}

Cross-validated free R value (termed R_{free}) was introduced to address the over-fitting problem in biomolecular crystallography[26,] and acts as an indicator of the fit between the experimental data and the refined structure, without the influence and bias of the refinement target itself throughout the process. Among our tests, all the final free R values obtained by DCN have been substantially improved over the conventional method, with a range from 0.0014 to 0.0675 (Table 3-2, Figure 3-8). Fourteen out of sixteen (87.5%) structures have been refined with an R_{free} improvement over 0.01. When compared with standalone DEN method, DCN achieves a 1.1x to 3.6x performance boost.

We illustrate the information for R_{free} (R_{free}-R_{work}, Ramachandran Statistics) in Figure 3-8 (Figure 3-9, Figure 3-10). For convenient comparison purpose, values for Conventional refinements are scaled to unity (other scaled accordingly), and values for DEN refinements are used to sort the PDB IDs on the horizontal axis. Pre-scaled data are taken from Table 3-2.

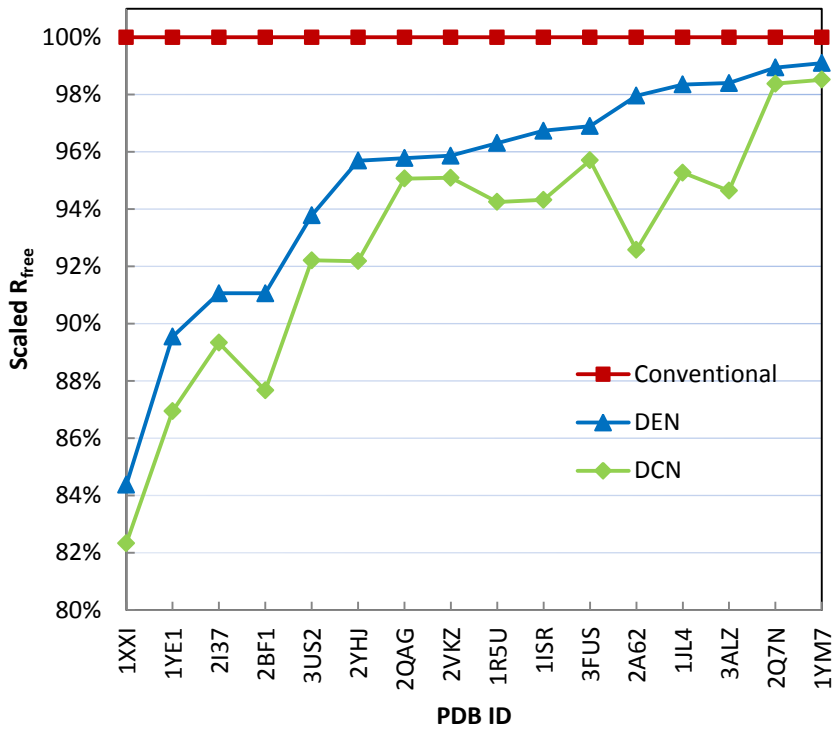


Figure 3-8 R_{free} of sixteen test systems for Conventional, DEN and DCN

3.3.2.3. Decrease in R_{free} - R_{work}

Degree of over-fitting could be assessed from the absolute value of difference between the free R and working R (termed R_{work}, the factor correlated with the maximum-likelihood scoring function and calculated using the reflections that are actually involved in the refinement process). Typically, R_{work} should be smaller than R_{free} due to the continuous optimization of maximum-likelihood function and high correlation between the function and R_{work}. In most of our test cases, DCN consistently delivers the smallest $R_{free} - R_{work}$ among all three methods (Table 3-2, Figure 3-9), thus minimizes the bias inherent in fitting the structure to the working set of reflection data throughout the

refinement process. The most favorable $R_{free} - R_{work}$ value for DCN is to the order of 10^{-3}, which almost eliminates the over-fitting effect, whereas for DEN and conventional, the best is to the 10^{-2}.

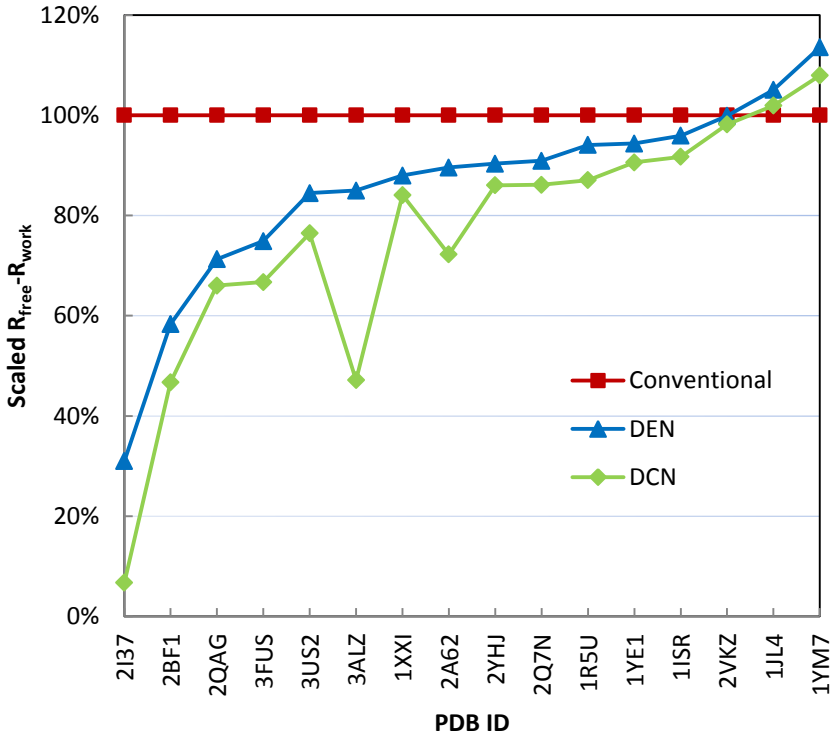


Figure 3-9 $R_{free} - R_{work}$ of sixteen test systems for Conventional, DEN and DCN

3.3.2.4. Increase in Ramachandran Statistics

To further evaluate the quality of the refined structures without the availability of a high-resolution model, we carried out the Molprobity structure validation[27]. The Ramachandran statistics was calculated to assess the quality of the secondary structures. Thirteen out of sixteen DCN-refined structures exhibit a larger percentage of residues that fall in the favored regions, resulting in a higher Ramachandran statistics, compared with

DEN-refined structures (Table 3-2,Figure 3-10). The restraints imposed by DCN add more geometry information from a high-resolution reference model, which usually possesses considerably accurate details on those secondary structures detected by the high-resolution diffraction data.

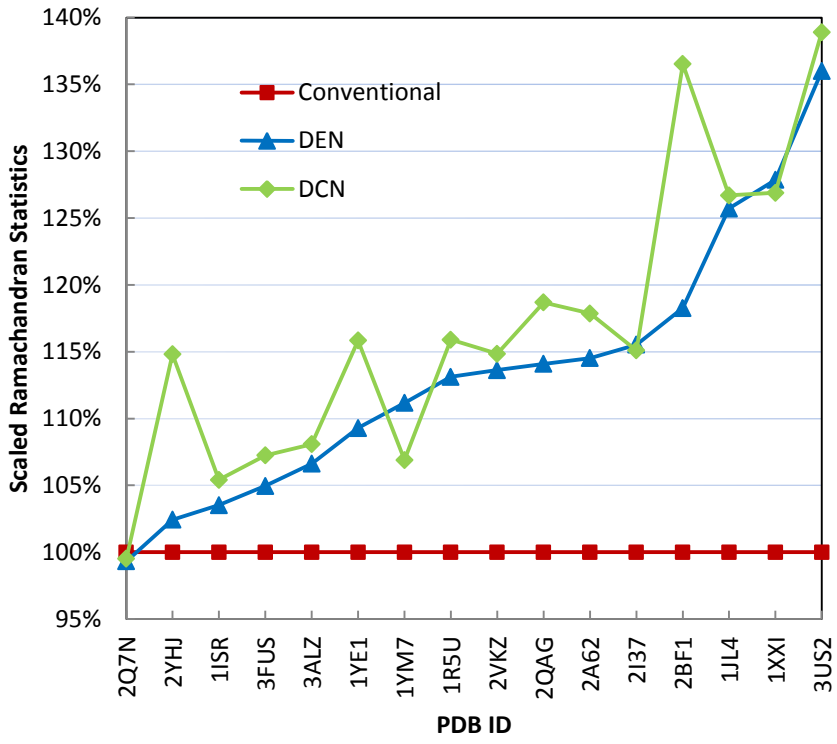
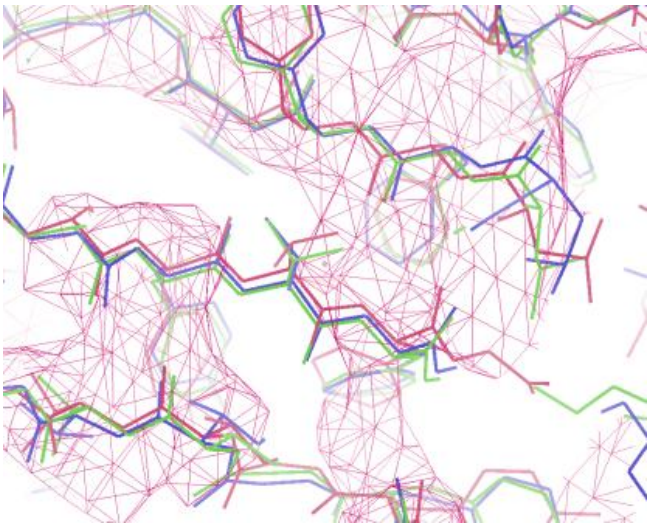


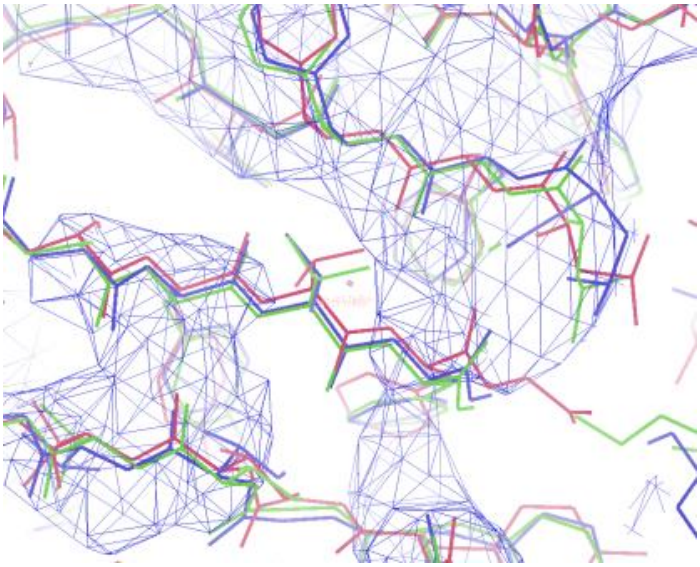
Figure 3-10 Ramachandran Statistics of sixteen test systems for Conventional, DEN and DCN

3.3.2.5. Improvement in electron density map interpretation

Along with the final structures, the phase combined sigma weighted $2F_o$-F_c electron density maps, derived from the experiment amplitudes and calculated model phases after the refinements with conventional (red), DEN (blue) and DCN (green) approach, have

been shown for different features. Case one (PDB ID 1JL4, Figure 3-11) illustrates the feature of DCN than improves the map and enhances the backbone interpretation. It is observed that, among all three maps, DCN (green) is the only one showing continuous main chain density. Clear breaks are, however, observed in both DEN and conventional maps. Case two (PDB ID 2BF1, Figure 3-12) shows the structure auto-correction feature of DCN. With a 0.04 improvement over conventional $R_{free,}$ DEN-refined structure allows remarkable residue shifts in several places on the main chain from the structure determined by conventional refinement. Nevertheless, it has been indicated by the corresponding density maps (blue mesh) that, selected shifts in DEN (e.g., C alpha atom in 368 GLY, chain A, shifts by 11.55Å and 9.83Å from itself in conventional and DCN, respectively) are poor-defined thus not reliable. DCN-refined structure is observed to be closer (2.24Å apart for C alpha) to the conventional-refined structure than DEN (9.83Å), and produces a self-consistent density map as well. Therefore, in some situations, DEN refined structure with significant branch deviations from conventional structure is not necessarily superior. DCN automatically reduces the remarkable disagreement and corrects the coordinates by favoring the better structure, and simultaneously fitting well to the density map generated by DCN itself. Even though appearing quite close to conventional structure, the DCN structure actually has a much higher quality with an improvement of 0.060 (Table 2-1) in R_{free} – that is more than 14% of its own value.





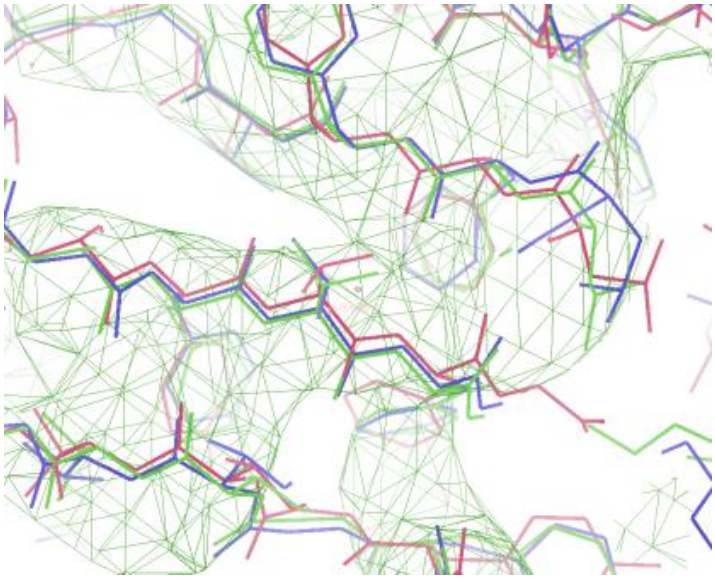
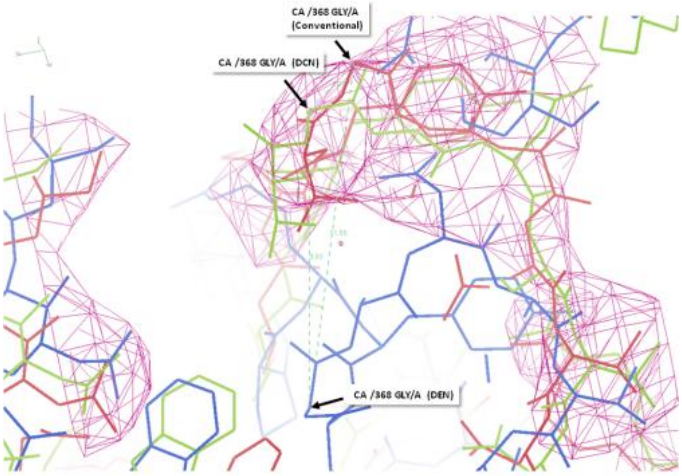


Figure 3-11 View of backbone trace. PDB ID 1JL4 centered on A23-THR is shown.



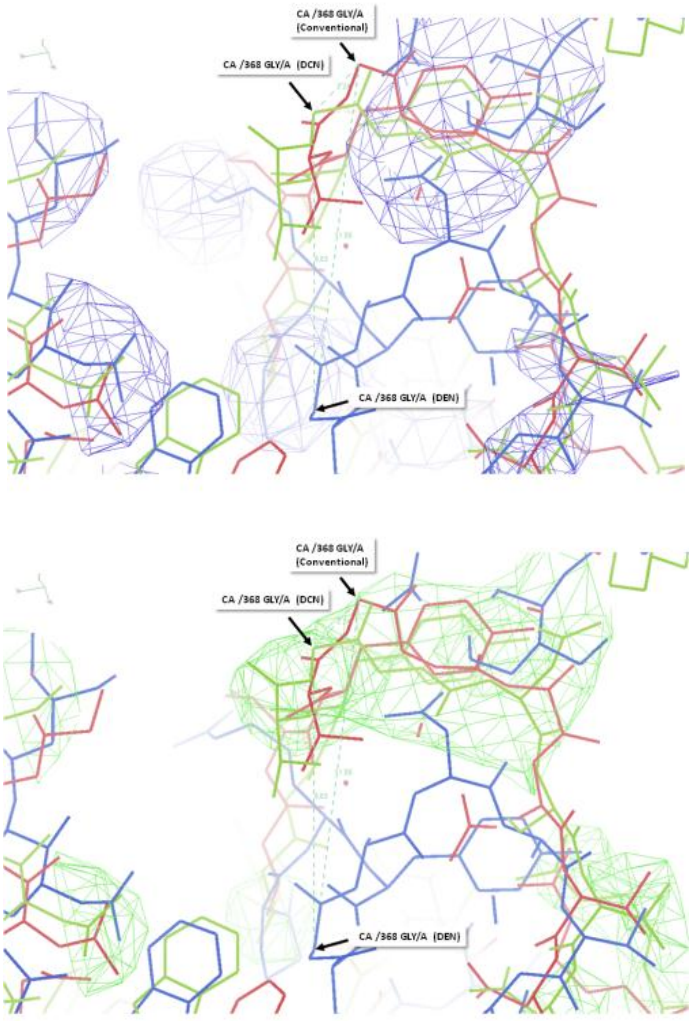


Figure 3-12 View of a remarkable branch deviation. PDB ID 2BF1 centered on A368-GLY is shown.

3.3.2.6. Re-refinement with NCS and experimental phase

DCN could easily incorporate other information to further facilitate the refinement process and improve the results (Table 3-3, Figure 3-13, Figure 3-14). We carried out refinements with Non-Crystallographic Symmetry (NCS) when related information was explicitly provided in the header of the PDB files. We then tested the effect of NCS information by intentionally turning NCS off before repeating the refinement. It is

observed that the NCS information has improved the structures for all three methods. Among them, DCN-refined structure is the most accurate in R_{free} no matter NCS is used or not. Therefore when applicable, it is encouraged to add NCS information during DCN refinement for seeking the best structure with the lowest R_{free}. In cases where experimental phase is obtained, for example, by single or multiple isomorphous replacement, using the MLHL[9] target function (with experimental phase) would produce better result than the MLF target (without experimental phase) for DCN in terms of Ramachandran statistics. Whereas conventional and DEN are also expected to benefit from the phase information, it is noted that, once again, improvement by DCN over the conventional and DEN refinements persists regardless of the availability of experiment phase data.

PDB ID	Target	NCS	R_{free}			DCN Improvement	Net gain fraction over DEN improvement	Ramachandran Statistics		
			Conventional	DEN	DCN			Conventional	DEN	DCN
1YM7	MLF	No	0.3398	0.3285	0.3169	0.0229	103%	0.596	0.683	0.640
	MLF	Yes	0.2764	0.2739	0.2723	0.0041	64%	0.703	0.781	0.751
3FUS	MLF	No	0.4584	0.4143	0.4001	0.0583	32%	0.373	0.550	0.533
	MLHL	No	0.4187	0.4057	0.4007	0.0159	38%	0.537	0.563	0.576

Table 3-3 Refinement with and without NCS or experiment phase information

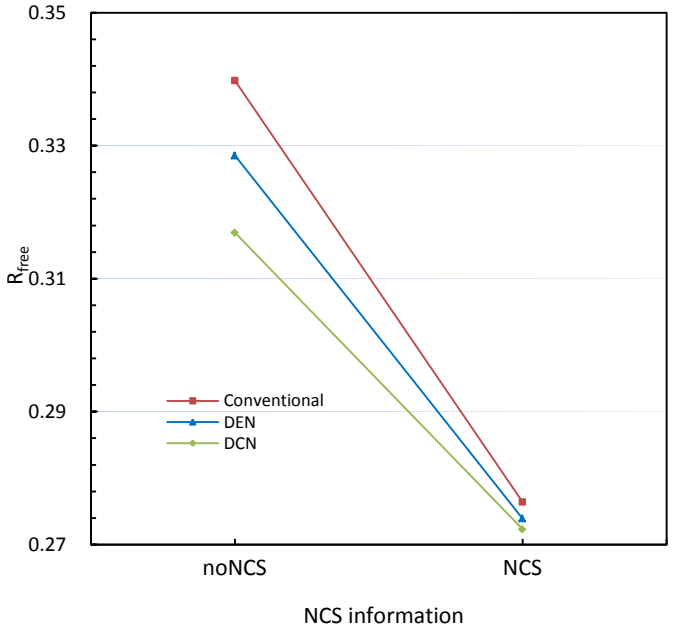


Figure 3-13 R_{free} vs availability of NCS information

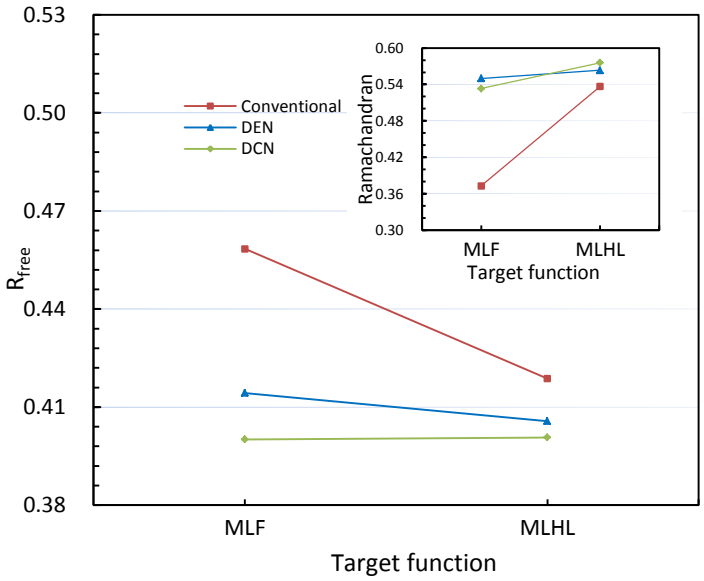


Figure 3-14 R_{free} (and Ramachandran) vs availability of experiment phase.

3.4. Supplementary Information

Table 3-4 A list of the structure property of all the re-refinement cases.

PDB ID	Resolution (Å)	Number of Chains	Sequence Length	No. of Observed Protein Residues	No. of All Observed Residues*	Ramachandran Statistics of Deposited Structure	Deposited R_{free}	Deposited R_{work}	Re-calculated R_{work}	Difference in R_{work}
1ISR	4.00	1	490	448	451	0.948	**0.259**	0.237	0.2414	-0.004
1JL4	4.30	4	557	557	557	0.922	*0.453*	*0.420*	0.3680	0.052
1R5U	4.50	11	4259	3517***	3527	0.805	0.373	0.345	0.2531	0.092
1XXI	4.10	10	3562	3532	3544	0.937	0.369	0.366	0.3929	-0.027
1YE1	4.50	4	574	574	772	0.968	0.343	0.295	0.2949	0.000
1YM7	4.50	4	2756	2422	2422	0.899	0.279	**0.224**	**0.2011**	0.023
2A62	4.50	1	322	319	325	0.749	0.346	0.271	0.2690	0.002
2BF1	4.00	1	316	304	354	*0.680*	0.388	0.385	*0.3920*	-0.007
2I37	4.15	3	1044**	954	975	0.896	0.382	0.377	0.3232	0.054
2Q7N	4.00	4	1336	1320	1365	0.793	0.287	0.237	0.2793	-0.042
2QAG	4.00	3	1206	702	705	0.895	0.392	0.376	0.3652	0.011
2VKZ	4.00	6	11814	10941	10947	0.935	0.268	0.268	0.2305	0.037
2YHJ	4.00	2	638	570	570	**0.977**	0.300	0.247	0.3150	*-0.068*
3ALZ	4.51	2	630	526	528	0.812	0.338	0.326	0.2276	**0.098**
3FUS	4.00	1	316	304	359	0.700	0.354	0.346	0.3775	-0.032
3US2	4.20	14	1624	1500	1584	0.886	0.334	0.326	0.3851	-0.059
Average	4.20	4.4	1965	1781	1812	0.863	0.342	0.315	0.3072	0.008
Maximum	4.51	11	11814	10941	10947	**0.977**	*0.453*	*0.420*	*0.3929*	**0.098**
Minimum	4.00	1	316	304	325	*0.680*	**0.268**	**0.224**	**0.2011**	*-0.068*

*All observed residues denote the sum of residue entries of protein, nucleic, heterogen, solvent that are observed and used in the refinement.

**Sequence length of 2I37 in PDB website is, however, recorded as 1047. This is because three modified residues of ACE were categorized as heterogen entries in the PDB structure file, but denoted as 'X' and included in the FASTA sequence file. These residues did not take part in the homology modeling process, did not have a corresponding residue in the reference structure, and as a result were not counted into sequence length or protein backbone residues.

***Chain M of 1R5U consists of unknown residues (UNK) and were excluded before refinement. Therefore number of protein residues is smaller than sequence length minus number of missing residues.

The sequence length varies from 316 to 11814 and represents a broad range of proteins with various sizes. Deposited values of R_{free} and R_{work} were directly taken from the PDB header. Re-calculated R_{work} denotes the value determined by CNS at the initial stage of refinement. Exact values could fluctuate because of multiple factors including but not limited to bulk solvent and B-factor correction, software and hardware environment of computing. The most and least favorable values in difference of R_{work} are 0.098 and -0.068, with an average difference of 0.008 (a good overall reproduction).

Moreover, the proportion of cases that results in a lower reproduced R_{work} is 9 out of 16 (56.25%). It is a substantial improvement over previous work[12] where these values are 0.053 and -0.109, with an average of -0.025 and only 4 out of 19 (21.05%) cases with lower reproduced R_{work}, respectively. This is due to the inclusion of all observed residues in the refinement together with other reasons stated above.

Table 3-5 A list of property of experiment data and reference model.

PDB ID	Resolution (Å)	Total No. of Diffractions		No. of Diffractions per Residue		Reference Model	
		Working	Free	Working	Free	Resolution (Å)	Sequence Identity
1ISR	4.00	6628	552	14.70	1.22	2.20	99.8%
2BF1	4.00	5842	280	16.50	0.79	1.99	*37.0%*
2Q7N	4.00	35237	1852	25.81	1.36	2.06	74.1%
2YHJ	4.00	11151	555	19.56	0.97	1.75	**100.0%**
2QAG	4.00	40358	2124	**57.25**	**3.01**	2.60	69.5%
3FUS	4.00	5841	279	16.27	0.78	2.20	39.1%
2VKZ	4.00	160231	8547	14.64	0.78	*3.10*	96.0%
1XXI	4.10	35818	4020	10.11	1.13	2.64	99.8%
2I37	4.15	11807	645	12.11	0.66	2.20	**100.0%**
3US2	4.20	14037	744	8.86	0.47	1.82	81.3%
1JL4	4.30	5880	645	10.56	1.16	2.25	94.7%
1YE1	4.50	3442	354	*4.46*	*0.46*	**1.43**	99.0%
2A62	4.50	3929	323	12.09	0.99	2.00	**100.0%**
1R5U	4.50	56023	1721	15.88	0.49	2.28	79.5%
1YM7	4.50	23110	1210	9.54	0.50	2.60	39.4%
3ALZ	4.51	13413	703	25.40	1.33	2.73	92.2%
Average	4.20	27047	1535	17.11	1.01	2.24	81.3%
Minimum	4.00	3442	279	*4.46*	*0.46*	**1.43**	*37.0%*
Maximum	4.51	160231	8547	**57.25**	**3.01**	*3.10*	**100.0%**

Diffraction data was fetched from the Protein Data Bank and converted into CNS recognized hkl file with no other modification. For data set without explicit R_{free} flag, a free data set was generated with the number of 5% of total diffractions by CCP4, a default value setting in the software. For several test systems (e.g. 1JL4), there exist diffraction entries with resolution higher than that given in the PDB header. Those entries were excluded before refinement and we only used the portion of data that agrees with the published resolution. Reference models were chosen according to several preferences

stated in ‘Method’. The sequence identity and resolution values in the table are linearly averaged by chain length according to the sequence, with information of resolution and identity of each chain’s corresponding template.

Table 3-6 Comparison of results between this work (ligands included) and previous work[12] (ligands excluded) with Conventional and DEN approach

PDB ID	Ligands not defined in CNS	Approach	R_{free} (this work)	R_{free} (previous work)[12]	Improvement
1ISR	GD	Conventional	0.223	0.237	0.014
		DEN	0.216	0.233	0.017
2BF1	BMA,NDG	Conventional	0.487	0.492	0.005
		DEN	0.443	0.479	0.036
1XXI	ADP	Conventional	0.382	0.465	0.083
		DEN	0.322	0.407	0.085
1YE1	HEM	Conventional	0.338	0.350	0.012
		DEN	0.302	0.312	0.010
2QAG	GTP,GDP	Conventional	0.405	0.401	-0.004
		DEN	0.388	0.392	0.004
2VKZ	CER,FMN	Conventional	0.312	0.337	0.025
		DEN	0.299	0.327	0.028
Average			0.343	0.369	0.026

Results of Conventional and DEN have been substantially improved in this work due to the inclusion of ligands during the refinement as well as factors such as software and hardware computing environment. This helps set up higher-standard controls in the first place. In most cases, improvement by DEN over Conventional is also larger in this work. DCN performance over DEN was calculated and compared with these ‘better’ DEN results.

3.5. Discussion and Implementation

More elaborate tailoring of DCN settings, such as carefully adjust the DCN model angle criteria, and select certain regions of molecule that have more reliable reference structures present, is expected to further enhance DCN's performance for structure predication, improvement and molecular-replacement phasing[37]. When best homology model found in the database does not have satisfactory sequence identity or resolution, it is possible to assign two distinct homology models to identical chain of a molecule for two parts of DCN, such that DAN and DEN information could be simultaneously absorbed from independent sources, to avoid the refinement from being guided under a single reference structure in an unfavorable or unreliable direction. Also, deformation of angular network and distance network do not need to be synchronous. A more robust conformation sampling may emerge when uneven frequencies or interleaved phases of the deformation period for the two networks are deployed. Moreover, DCN could be effortlessly implemented in grid computing servers with an online GUI[38], allowing interested users to carry it out via a web portal with ease.

References

1 Figures obtained from Wikimedia Commons, a freely licensed media file repository. *commons.wikmedia.org*.

2 PDB File Format. *Contents Guide Version 3.30* (2011).

3 Urzhumtsev, A. G. & Podjarny, A. D. On the problem of solvent modelling in macromolecular crystals using diffraction data: 1. The Low Resolution Range. *CCP4 Newsletter* **31**, 12-16 (1995).

4 http://www.csb.yale.edu/userguides/datamanip/xplor/xplorman/node284.html. *Xplor Online Manual*.

5 Read, R. J. & Moult, J. Fitting electron density by systematic search. *Acta Cryst. A* **48**, 104-113 (1992).

6 Ten Eyck, L. F. & Watenpaugh, K. D. *International Table for Crystallography*. **F**, 369.

7 Bricogne, G. & Gilmore, C. J. A multisolution method of phase determination by combined maximization of entropy and likelihood. I. Theory, algorithms and strategy. *Acta Cryst. A* **46**, 284-297 (1990).

8 Pannu, N. S. & Read, R. J. Improved structure refinement through maximum likelihood. *Acta Cryst. A* **52**, 659-668 (1996).

9 Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Cryst. D* **54**, 1285-1294 (1998).

10 Engh, R. & Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst. A* **47**, 392-400 (1991).

11 http://www.csb.yale.edu/userguides/datamanip/xplor/xplorman/node183.html. *Xplor Online Manual*.

12 Schroder, G. F., Levitt, M. & Brunger, A. T. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **464**, 1218-1222 (2010).

13 Ramachandran, G. N., Ramakrishman, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95-99 (1963).

14 Brunger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-475 (1992).

15 Stout, G. H. & Jensen, L. H. *X-ray structure determination, a partial guide 2nd edn*, 343-378 (1989).

16 Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* **32**, 922-923 (1976).

17 Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* **34**, 827-828 (1978).

18 Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**, 3370-3374 (2003).

19 Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710 (2004).

20 Levitt, M. & Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA* **95**, 5913-5920 (1998).

21 Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16**, 776-785 (2000).

22 Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. & Elofsson, A. A study of quality measures for protein threading models. *BMC Bioinformatics* **2**, 5 (2001).

23 Rice, L. & Brunger, A. T. Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* **19**, 277-290 (1994).

24 Go, N., Noguti, T. & Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**, 3696-3700 (1983).

25 Schroder, G. F., Brunger, A. T. & Levitt, M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15**, 1630-1641 (2007).

26 Terwiliger, T. C. & Eisenberg, D. Isomorphous replacement: effect of errors on the phase probability distribution. *Acta Cryst. A* **43**, 6-13 (1987).

27 Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res* **32**, W615-W619 (2004).

28 Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671-680 (1983).

29 Hendrickson, W. A. & E.E, L. Representation of phase probability distributions for simplified combination of independent phase information. *Acta Cryst. B* **26**, 136-143 (1970).

30 Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259-264 (2007).

31 Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815 (1993).

32 McCoy, A. K. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658-674 (2007).

33 Brunger, A. T. *et al.* Crystallography & NMR System (CNS), a new software suite for macromolecular structure determination. *Acta Cryst. D* **54**, 905-921 (1998).

34 Brunger, A. T. Version 1.2 of the Crystallography and NMR System. *Nature Protocols* **2** (2007).

35 Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **14**, 33-38 (1996).

36 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and Development of Coot. *Acta Cryst. D* **66** (2010).

37 Brunger, A. T. *et al.* Application of DEN refinement and automated model building to a difficult case of molecular replacement phasing: the structure of a putative succinyl-diaminopimelate desuccinylase from Corynebacterium glutamicum. *Acta Cryst. D* **68**, 391-403 (2012).

38 O' Donowan, D. J. *et al.* A grid-enabled web service for low-resolution crystal structure refinement. *Acta Cryst. D* **68**, 261-267 (2012).